Simple and rationale-providing SMS reminders to promote accelerometer use: A within-trial randomised trial comparing persuasive messages

Matti T. J. Heino[1, 2*], Ari Haukkala[2], Keegan Knittle[2], Tommi Vasankari[3], Nelli Hankonen[1, 2]

[1] Department of Social Sciences, University of Helsinki, Helsinki, Finland;

[2] Department of Social Sciences, University of Tampere, Tampere, Finland

[3] UKK Institute for Health Promotion Research, Tampere, Finland

* Corresponding author

E-mail: matti.tj.heino@gmail.com

Abstract

Background: Literature on persuasion suggests compliance increases when requests are accompanied with a reason (i.e. the "because-heuristic"). The reliability of outcomes in physical activity research is dependent on sufficient accelerometer wear-time. This study tested whether SMS reminders—especially those that provided a rationale—are associated with increased accelerometer wear-time.

Methods: We conducted a within-trial partially randomised controlled trial during baseline data collection in a school-based physical activity intervention trial. Of 375 participants (mean age=18.1), 280 (75%) opted to receive daily SMS reminders to wear their accelerometers. These 280 participants were then randomised to receive either succinct reminders or reminders including a rationale. Data was analyzed across groups using both frequentist and Bayesian methods.

Results: No differences in total accelerometer wear minutes were detected between the succinct reminder group (Mdn=4909, IQR=3429–5857) and the rationale group (Mdn=4808, IQR=3571–5743); W=8860, p=0.65, CI95=-280.90–447.20. Similarly, we found no differences in wear time between participants receiving SMS reminders (Mdn=4859, IQR=3527–5808) and those not receiving them (Mdn=5067, IQR=3201–5885); W=10642.5, p=0.77, CI95=-424.20–305.30. Bayesian ANOVA favored a model of equal weartime means, over one of unequal means, by a Bayes Factor of 12.05. Accumulated days of valid accelerometer wear data did not differ either.

Equivalence testing indicated rejection of effects more extreme than a Cohen's d (standardised mean difference) of ± ~0.3.

Conclusions: This study casts doubt on the effectiveness of using the because-heuristic via SMS messaging, to promote accelerometer wear time among youth. The because-heuristic might be limited to face-to-face communication and situations where no intention for or commitment to the behavior has yet been made. Other explanations for null effects include non-reading of messages, and reminder messages undermining the self-reminding strategies which would occur naturally in the absence of reminders.

Trial registration: DRKS DRKS00007721. Registered 14.04.2015. Retrospectively registered.

Keywords: Accelerometry, intervention, text messaging, SMS, persuasion, adherence, behaviour change, adolescents, school-based research, partially randomised trial

Background

Compliance with accelerometer wear instructions

Reliable and valid assessment is necessary when evaluating whether policies or interventions change physical activity (PA) levels in the target group. Little consensus exists about what to measure, when, with what and for how long in PA research [1,2]. While an inability of individuals to accurately remember their past PA and social desirability are clear problems with self-reported PA measures [3], objective measurements of PA (e.g. pedometers and accelerometers) have issues too. Zhuang et al. [4] found that missing accelerometry data was more common in 15- to 17-year-olds than among younger participants, especially during weekends (Sundays in particular), with missing data occurring increasingly from the first recording day to the last. This exemplifies a key issue in measurement: the proportion of an individual's day or week captured by the measure. An extreme example would be an individual, who only wears the measurement device when undertaking PA. Thus, some guidelines suggest that a person should wear an accelerometer for a minimum of 10 hours daily for at least 4 days in a 7-day measurement period in order to obtain an accurate reading of PA [1,2]. Participants' compliance with instructions on wearing the accelerometer is clearly very important in obtaining accurate PA measurements [5].

Research on enhancing accelerometer instruction compliance rates is rare [2,6], particularly among older adolescents. One strategy has been monetary incentives contingent on proper wear-time [7]. Sallis et al. [8] used an alternative strategy,

asking participants to re-wear the accelerometer if they had not worn it for at least 5 valid days (>10 valid hours of data) or a minimum of 66 valid hours across 7 days.

Barak et al. [9] suggest that new opportunities to promote compliance—such as text messaging—may be more reliable and effective than traditional methods, such as written or verbal wear instructions by the investigator. Zhuang et al. [4], too, recommend SMS reminders. Toftager et al. [10] used SMS reminders to increase compliance but did not report effects or acceptability. In a self-selected Irish sample of adolescents [11], daily SMS reminders were associated with putting on the accelerometer in the morning, but not in increased overall compliance (defined as valid days of data or minutes of non-wear). The study did not report levels of wear or effects of the reminders. The discrepancy between remembering to put on the device and actually wearing it for a sufficient amount of time indicates that these may be separate behaviors.

Compliance and the 'because-heuristic'

Since the classic "Xerox machine study" by Langer, Blank and Chanowitz (henceforth: *LBC*) [12], providing reasons for compliance has been discussed in the social influence literature. The study indicated that placebic or pseudo-reasons [13] ("*Excuse me, I have 5 pages. May I use the xerox machine, because I have to make copies?*"; 93% compliance) could result in similar compliance rates as actual reasons ("[...] *because I'm in a rush?*"; 94% compliance) compared to the request only condition ("*Excuse me, I have 5 pages. May I use the xerox machine*"; 60% compliance). Pratkanis (2007), identified "placebic reasons" in his index of social influence tactics, but called for further research into the subject. Less careful are Cialdini, Goldstein and Martin [14], who tout the "unique motivational influence of

the word *because*", basing their claims on the importance of reasoning in social influence. To this day, the xerox machine study remains cited in the press as an example of the power of the word 'because' [15–18].

> A well-known principle of human behavior says that when we ask someone to do us a favor we will be more successful if we provide a reason. People simply like to have reasons for what they do. [19]

Following the terminology used by Key, Edlund, Sagaring and Bizer [20], the phenomenon of increased compliance by providing reasons is referred to as "the because-heuristic." Let us accordingly define the *naïve because-heuristic* as "reasons increase compliance."

In the LBC study 1, this effect of reasons increasing compliance was only found when the favor asked for was small (five instead of ten pages, translating to effect sizes of d=0.87 and d=0.13, respectively) [12]. Still, the results in general, as well as their implications have been questioned [21,22]. A study by Folkes suggests, that instead of the size of the request, the effect is moderated by controllability [21]. Pooling Folkes' reason conditions results to a d=-0.026, speaking against the quote above, and pointing out that the "power of reasons" effect is malleable, in the least.

To our knowledge, only one published direct replication of the LBC study 1 exists [20]. The main effect of the study replicated (d=0.67 for placebic over no reason and d=0.69 for real over no reason conditions), although over 20% (34 out of 163) of the participants needed to be excluded for various reasons. Lack of published replication studies, of course, is not new in the field of psychology [23].

In a conceptual replication of the phenomenon, in small request conditions, reasons (either placebic or real) increased compliance by an equivalent of d=0.43 (calculated from Table 1 of [24]) when including their additional persuasion group and d=0.22

when excluding it. Another conceptual replication [25] found d=0.15 for requests perceived as small, and d=0.21 for requests perceived as large (as calculated from figure 3 of [25]).

These studies seem to temper earlier claims for the power of reasons in increasing compliance. In contrast to the *naïve because-heuristic*, let us define the *weak because-heuristic* as "reasons increase compliance, but only if the perceived favour is small".

The Let's Move It cluster randomized trial

Inadequate PA predicts increased morbidity and mortality in people of low socioeconomic status (SES) [26], with SES differences in PA emerging already in adolescence [27]. Finnish vocational school students are less physically active than those in high school [28]. The Let's Move It intervention aimed to increase PA and decrease sedentary behaviors in older adolescents in vocational schools.

The current study was conducted as a sub-study of the cluster randomised effectiveness evaluation trial of the Let's Move It intervention [29]. In a preceding feasibility study [30], participants' accelerometer wear times were suboptimal; 47% (18/38) of baseline participants reached the cutoff of ten hours per day for at least four days, 63% (17/27) for the first and 75% (9/12) for the second follow-up. A frequently cited explanation for not wearing the accelerometer was forgetting to put on the device.

Aims and hypotheses

In this within-trial study, we investigate SMS-reminder strategies to improve the duration of accelerometer wear time. The literature cited above lead us to hypothesise that reminders would increase accelerometer wear time and that citing reasons would amplify the effect. In addition to daily wear hours, we are interested in the number of days our participants provide valid activity data (i.e. days of ≥ 10 hours of activity data). The target behavior is thus twofold: 1) putting on the accelerometer in the morning for as many days as possible, 2) wearing the accelerometer for as long as possible in the waking hours each day. In this study, two main research questions are posited:

1. Are SMS-reminders associated with greater accelerometer wear times?

The current study investigated this by comparing the compliance rates across a) participants who opted to receive SMS reminders to wear their accelerometer, and b) participants who opted not to receive the reminders (non-randomised control group).

Substantive hypothesis S_1: If forgetting is an important reason for non-compliance, in the absence of intervening factors, reminders should increase compliance.

Statistical hypothesis H_1: Those who receive SMS reminders will have higher accelerometer wear times than those who do not.

2. Does offering reasons to comply affect accelerometer wear time?

Substantive hypothesis S_2: If reasons increase compliance, SMS reminders containing reasons to wear an accelerometer should lead to greater compliance.

Statistical hypothesis H_2: Those who receive reasons in the SMS reminders have more minutes of accelerometer wear and more days of valid data (≥10 hours of activity) than those who do not receive reminders containing a reason.

An additional research question, on whether providing reasons to comply with accelerometer wear increases trial retention, is omitted here. These null results are reported in [31].

Methods

The design of this study was a within-trial, outcome-assessor blinded, partially randomised controlled trial (RCT). In addition to the randomised experiment between two message types, quasi-experimental data were acquired from a self-selected opt-out arm. This study was conducted during the baseline assessment of the first two recruitment waves (out of six; the internal pilot study) of the Let's Move It cluster-randomised controlled trial [29]. This article is based on unpublished work available at https://osf.io/89mhu/. Additional information on methods and results, in addition to all analysis code, can be found in the supplementary website at https://git.io/vNl8X (permalink provided in [32]).

Participants and sampling procedures

To be included in the study, the participants had to fulfill inclusion criteria of the Let's Move It study [29] and had to have consented to the accelerometry measurements: all were at least 16 years old and were vocational school students. The reminder arms consisted of the participants who opted in to receive reminders for accelerometer wear.

During baseline recruitment of the Let's Move It trial internal pilot study, students in two vocational schools were approached during class and informed about their school's study participation in the study. After the invitation to participate in the main trial and collection of signed informed consent forms, those who consented were given an online questionnaire to complete. Details of trial procedures are reported in the protocol [29].

After 1-3 days, research assistants gave the participants a waist-worn accelerometer (Hookie AM 20, Traxmeet Ltd, Espoo, Finland) and instructed them on how to wear it for a duration of seven consecutive days (including the day of receiving the device). When participants received the accelerometers, they were asked whether they would like to receive SMS messages to help them remember to put it on every morning. Those who consented to the messages were subsequently randomised to one of two message conditions, and those who opted not to receive the reminders were treated as a self-selected control arm.

After seven days, participants returned their devices to research assistants and were asked to fill out a short questionnaire assessing process measures (see Appendix 3).

Random assignment

Participants were assigned to the reason and succinct arms after they were recruited. The first author extracted the phone numbers from the list and used R code to create an amount of random numbers equal to the number of new participants. The vector of random numbers was then assigned to the participants. Participants with a number equal to or smaller than the median of the vector were allocated to the reason-condition. Others were allocated to the succinct condition. Research assistants

working in the field were blind to group allocation Recruitment and randomisation took place on the same day, and restrictions such as blocking or stratification were not used.

Recruitment took place in two waves, alongside the recruitment of the main trial. In order to increase the rates of participants opting in for the reminders, the recruitment prompt was slightly modified for the second wave. The research assistants presented the SMS reminders as the default option, and asked whether this is acceptable to the participants.

Random assignment was not visible to the participants and the research assistants did not mention that different kinds of messages were going to be sent. The statistician who analysed the raw accelerometer data was blind to group assignment.

Interventions

An important issue regarding the current study was to avoid tampering with the effects of the main trial. In other words, it should not affect main trial outcome measures in any other ways except for increased data quality. Care was taken to formulate the SMS messages to not pressure participants or provoke changes in main trial outcome measures such as PA.

We altered a previous procedure [11] by varying the message content slightly each day to reduce habituation and thus expected to increase the chances of the message being read, for both arms.

The two arms received different message content.

a) **Succinct reminder condition:** 1. a greeting – 2. a reminder – 3. a thank you

b) **Reminder and reason:** 1. a greeting – 2. a reason beginning with "Because…", followed up with a reminder – 3. a thank you

Messages are presented in detail in Table 1 below.

Table 1. SMS content, translated to English.

Morning	Reminder with rationale (the "because heuristic")	Succinct reminder
1st	Morning! Because your participation is precious, please remember to put on the motion measurement device and wear it until you go to sleep (except in the shower etc.) - thanks!	Morning! This is a reminder to put on the motion measurement device and wear it until you go to sleep (except in the shower etc.) - thanks!
2nd	Hi! Because you're aboard in producing very important knowledge, please remember to put on the motion measurement device now and wear it as instructed until you go to sleep. Thanks a lot!	Hi! Please remember to put on the motion measurement device now and wear it as instructed until you go to sleep. Thanks a lot!
3rd	Hello! Because the study wouldn't succeed without your help, please remember to put on the motion measurement device again and wear it until you go to sleep (except in the shower etc.) - thanks!	Hello! Please remember to put on the motion measurement device again and wear it until you go to sleep (except in the shower etc.) - thanks!
4th	Morning! Because the data you gather is highly valued, please remember to put on the motion measurement device and wear it until you go to sleep. Thanks (we're already past midpoint)!	Morning! Please remember to put on the motion measurement device and wear it until you go to sleep. Thanks (we're already past midpoint)!
5th	Howdy! Because your participation produces very important knowledge, please remember to put on the motion measurement device and wear it until you go to sleep (except in the shower etc.) - thanks!	Howdy! Please remember to put on the motion measurement device and wear it until you go to sleep (except in the shower etc.) - thanks!

6th	Hi! Because even this last day is important, please remember to put on the motion measurement device and wear it until you go to sleep. Return the motion measurement device to school tomorrow - thanks!	Hi! Please remember, even on this last day, to put on the motion measurement device and wear it until you go to sleep. Return the motion measurement device to school tomorrow - thanks!

We sent the messages using an SMS Gateway device MT-SF100-G-EU (MultiModem iSMS Server 1-port) by Multi-Tech Systems (http://www.multitech.com/brands/multimodem-isms). We used a manufacturer-designed guided user interface for the first recruitment wave and a custom interface designed by a local service provider for the second wave.

Registration and deviations from registered plan

The study plan was reviewed by the Ethics Committee for Gynaecology and Obstetrics, Pediatrics and Psychiatry of the Hospital District of Helsinki and Uusimaa (decision number 367/13/03/03/2014).

Official public registration in the German Clinical Trials Register (DRKS-ID: DRKS00007721) was completed three months after recruitment of the first wave had been initiated, but before data was available. Pre-registration (before starting data collection) failed due to lack of available resources at the time.

The original plan was to establish the additive effect of messages containing a reason and those not containing one over a no-message condition during the baseline measurement of the first batch. With the sample size we expected (n=140), we would have had over 95% power to detect an effect of d=0.6 (slightly smaller than the one discovered in the LBC replication [20]). We had then planned to pit the more successful message type against a third message in the second wave. Instead of going forward with the plan of using a third message, we made the decision to gather

another wave of participants with the same message types after the data from the first wave was analysed. This was due to the fact that, contrary to our expectations, no difference between the two messages was detected. This is important to note, as it means we can no longer rely on a long-term error rate of 5% [33] and—as p-values depend on the sampling distribution—default p-values from common statistical programs no longer apply [34].

To address the issue of inadequate reporting in the sciences [35], the current report complies with the Consolidated Standards of Reporting Trials (CONSORT) statement [36]. Contributor roles are clarified in Appendix 1, according to a taxonomy for this purpose [37].

Outcomes

Primary outcome measures

Primary outcome measures were 1) accelerometer wear time minutes and 2) days with ≥ 10 hours of valid accelerometer data. As this trial was conducted within a larger trial, several other measures were collected and are listed in the Let's Move It protocol [29]. The main trial used a 3-axis accelerometer with a 2GB internal memory (Hookie Meter v2.0, Hookie Technologies Ltd, Espoo, Finland). The activity data was registered using raw data and a 100 Hz sampling rate.

Implementation assessment measures

A one-page questionnaire (Appendix 3) was used to gain additional insight into the reception of the messages.

Self-reported message receipt. As we could not gather objective log data on the number of messages opened, we asked participants to assess on how many mornings they had opened and read the SMS. Response options were: Not on a single morning, On 1 morning, On 2–3 mornings, On 4–5 mornings and Every morning.

Manipulation and contamination check. As participants were randomised individually, as opposed to clusters at school class level, discussing the SMS messages with their classmates could have led to students finding out that not everyone received the same messages, and perhaps also reveal the study hypotheses. We attempted to gauge the extent of this by asking them how often they had discussed the messages with peers. Response options were Not once, Once, 2–3 times, 4–5 times and More often.

Acceptability of SMS message content was assessed by asking the participants, how much they agree with the statement "I was satisfied with the content of the messages". Response options again had a 5-point scale: Completely disagree, Somewhat disagree, Do not agree nor disagree, Somewhat agree and Completely agree.

Statistical analyses

All non-Bayesian analyses were conducted using RStudio running R [38,39]. Plots were drawn using R packages 'ggplot2' [40] and 'yarrr' [41]. Distributions between

the reason and succinct groups in the implementation assessment questions were compared using the chi-square test.

Accelometer wear times were analysed using bootstrapping methods. A 95% bootstrap confidence interval for a mean can be acquired by resampling observed data to simulate a sampling distribution, obtaining the values for the 0.025^{th} and 0.975^{th} percentiles of resampled means [42]. A kernel density plot, bootstrap confidence interval and a bootstrap test of equivalence were conducted using R package 'sm' [43] for differences of distributions of the two reminder arms. Wilcoxon rank sum test with continuity correction was used to compare medians between groups.

ANOVA for equivalence of means between the two reminder groups and the no-reminder group, as well as its illustration, was performed using R package 'userfriendlyscience' [44]. Additionally, a MANOVA with wear time minutes and wear days with valid data as dependent variables, and SMS group as an independent variable, was used to test robustness of results.

A 95% Bayesian Highest Density Interval (HDI) [41] of the means of valid wear days was plotted using R package 'yarrr'. HDI refers to the most likely population parameter values (here: means) given the data; information which is not delivered by frequentist confidence intervals [45,46].

Bayes Factors. Due to our sampling methods (e.g. decision to collect more data was based on observed data), traditional frequentist statistics faced limitations. Thus, we also calculated Bayes Factors [47–49] for our main outcome measures. A Bayes factor BF_{01} is essentially the ratio of two likelihoods, answering questions such as "Given the data, how many times more likely is the null hypothesis, compared to a

specific alternative hypothesis". We used the R package BayesFactor [50]; For comparing means, this package assigns the alternative hypothesis a Cauchy prior. We used a prior scale of 0.3, in accordance with common effects in health psychological research [51]. This reflects a prior belief that 50% of the effects lie between d=-0.3 and 0.3. For contingency tables, priors are described in Jamil et al. [52]. The minimum value is 1, and an increase reflects the belief, that the distribution of observations in the given categories under H1 is relatively more similar to H0. Additional information on inference using Bayes Factors, and prior robustness checks are found in the supplementary website.

Equivalence testing. In the frequentist statistical paradigm, support for the null hypothesis is indicated by the practice of equivalence testing [53]. For a difference between means, one essentially first establishes a region of equivalence to zero, then conducts and combines two t-tests. The first one tests whether the effect is higher than the lower bound (in our case, -0.3), and the other tests whether the effect is smaller than the higher bound (in our case, 0.3). The tests were conducted using R package "TOSTER" [54].

We did not conduct multi-level analyses to account for the intra-class correlation of 0.09 for total accelerometer wear time. Heterogeneity analysis is presented in the supplementary website file under "Heterogeneity among clusters".

Using standard deviations estimated from feasibility study [30] data, we determined a practically significant effect size for wear time hours to be d=0.42 – enough to bring a person from 9.5 hours of daily data to reach the cutoff of 10 hours. For our purposes, we decided to consider effect sizes between -0.3 and 0.3 as equivalent to

zero. Additional details are presented in the supplementary website under “Statistical power”.

Analysis regarding statistical power is presented in Figure 1, holding alpha constant at 0.05 and sample size at achieved levels. As seen from the figure, we had 90% power to discover an effect of size d=0.39, 80% to detect d=0.3, 60% to detect d=0.27 and 40% to discover an effect of d=0.21. Thus, type 2 error probabilities were small for effects near our defined minimal effect size of interest, but high for small effects.

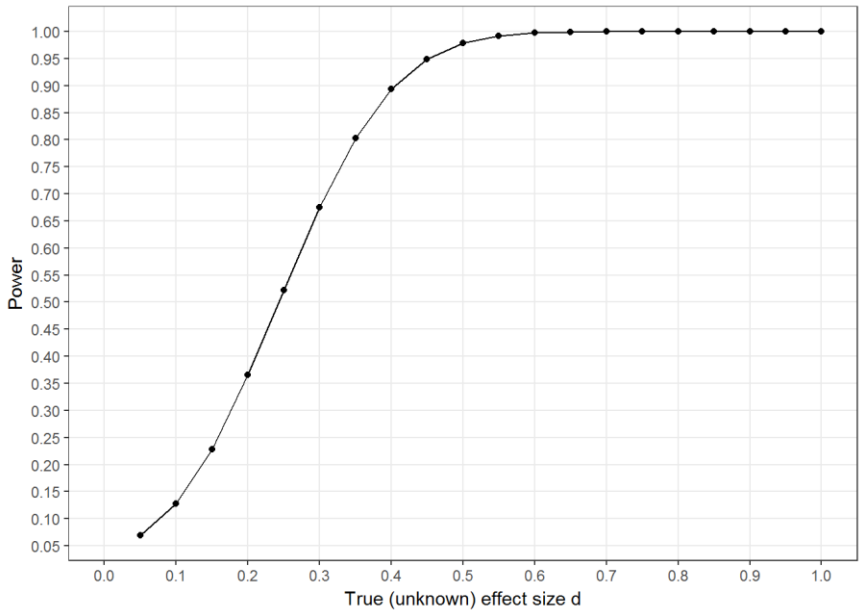


Figure 1: Statistical power, t-test for an unknown real effect.

We also evaluated Type S and type M error probabilities [55], and the v-statistic [56]. The analysis is presented in the supplementary website. In brief; our design was relatively well-equipped to handle medium-sized effects, but is subject to considerable bias under small effects.

Results

Descriptive data

A participant flow diagram presented in Appendix 2 indicates how the messages were sent to almost all participants as intended.

Of the 375 participants consenting to accelerometer measurements as part of the main trial, 95 opted out of receiving reminders and an additional 7 did not receive messages due to technical difficulties. In the end, the SMS messages with reasons were sent to 138 and the succinct messages to 135 participants. Consent rate for reminders was 54% (101 out of 186) in the first wave and 95% for the second wave (179 out of 189).

Table 2 shows the sample characteristics for the baseline data.

Table 2: Sample characteristics.

	SMS group				
	Reason	**Succinct**	**Opt out**	**Send failed**	**Total**
Total n	**138**	**135**	**95**	**7**	**375**
Weartime data available	133	129	83	7	352
Female	28 %	30 %	27 %	43 %	30 %
M age (SD)	17.9 (1.8)	18.2 (2.6)	18.9 (4.3)	18 (1.4)	18.3 (2.9)

Note: One person from both SMS groups missed the first message due to phone number imputation failure. This was considered to be of no practical consequence and they were counted as having received their intervention as planned.

Implementation and process measures

Manipulation and contamination check

The distributions in answers to opening and reading the SMS, discussing the messages with peers or being satisfied with the messages did not differ among the reason and succinct groups ($\chi^2(4)$=1.356, 2.566 and 3.903 respectively; all p's > 0.4). All Bayes Factors indicated strong support for the null (see supplement).

As shown in Figure 2, 74.9% of respondents reported having opened and read the SMS at least four mornings.

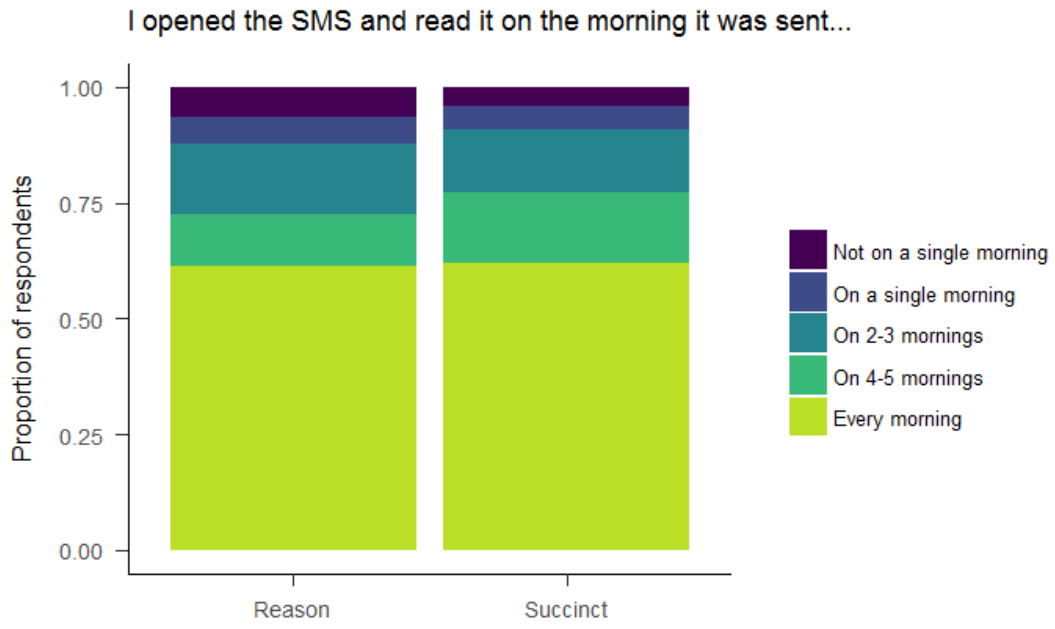


Figure 2: Opening and reading the SMS.

Discussing the content of the messages with peers was not common; 91.1% answered having done so never or just once (Figure 3).

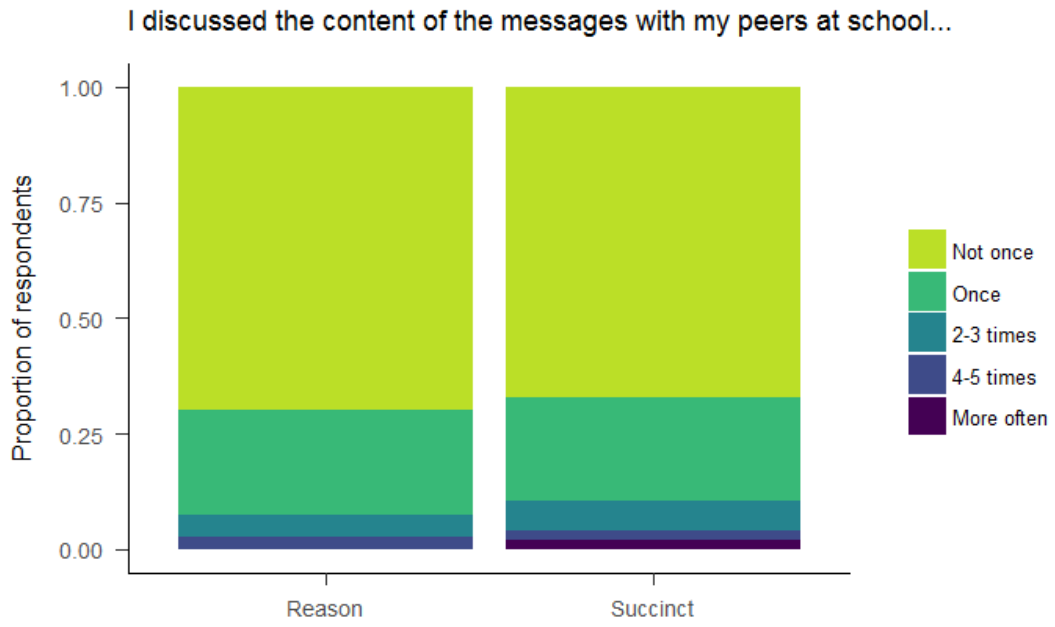


Figure 3: Discussing the SMS with peers.

Satisfaction with the message content and open comments

The messages were evaluated positively. Only 3.5% of the participants indicated disagreement with the statement “I was satisfied with the content of the messages” (see supplement).

Open comments did not reveal unforeseen negative effects. In addition, 13% (9 out of 70) of participants who answered the question explicitly added, that remembering to wear the device was due to receiving the messages.

Wear times

Wear time minutes

Accelerometer wear times did not indicate meaningful differences between groups (Figure 4).

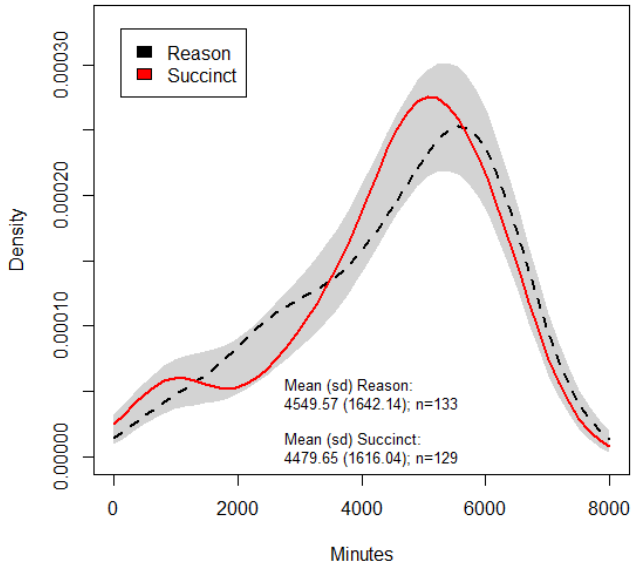


Figure 4: Total wear time in minutes (dashed line for the reason condition, solid for succinct). Grey band around the kernel density plots refers to 95% likelihood of containing the true density plot, if the two lines were generated by data from the same distribution.

Bootstrap tests of equal densities indicated no differences in total wear time minutes between the two message types (p=0.28), nor between those who received and did not receive messages (p=0.35). Wilcoxon rank sum test showed no differences in distributions between message groups (W=8860, p=0.647, CI95=-280.90–447.20) or whether one opted in the messages or not (W=10642.5, p=0.771, CI95=-424.20–305.30). Differences were neither detected between the two schools (W=17398.5,

p=0.051, CI95=-1.60–619.60) or recruitment waves (W=17310.5, p=0.067, CI95=-19.0–586.3).

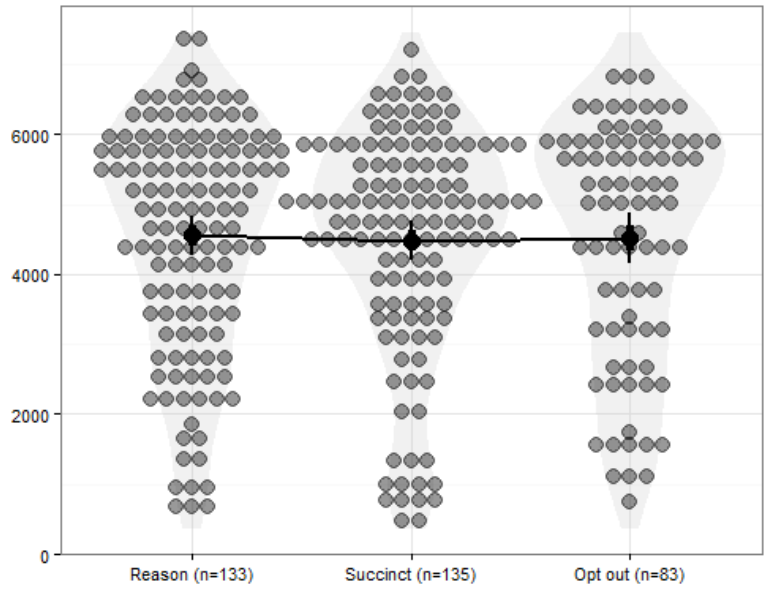


Figure 5: Means and the total wear time distributions of the three groups. Error bars indicate 95% confidence intervals. No differences are detected.

The violin plots above illustrate how wear times in all three groups are distributed. Bayesian ANOVA gives us BF_{01}=12.05, indicating strong evidence for equivalent means, against a model where all means are unequal. Prior robustness graph (see supplement) starting from r=0 depicted a convex function, where BF_{01} rises to 10 at r=0.27 and reaches 422.34 at r=2.00. Furthermore, BF_{01} relative to an ordered model of Reason > Succinct > Opt out was 23.07 (see section “Interpreting Bayes Factors” in the supplement).

Equivalence tests indicated, that the mean wear time differences between message types (69.92 minutes, 90% CI [-262.37; 402.21]) and the reminder/opt out groups (1.98 minutes, 90% CI [-347.12; 351.08]) were statistically significantly larger than d=-0.3 and smaller than d=0.3. In other words, the effect size for the difference in means was deemed less than |0.3|.

Valid measurement days

Figure 6 shows densities and spread of valid measurement days by group. As can be visually inspected from the HDIs, population means are equivalent.

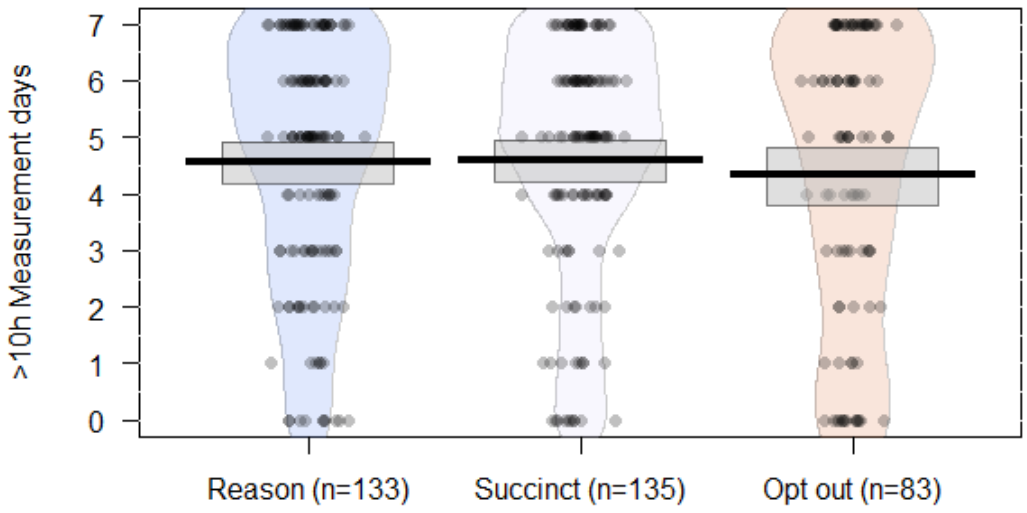


Figure 6: Measurement days of >10 hours of data gathered by group. Horizontal lines represent means, boxes Bayesian 95% Highest Density Intervals (with flat priors).

Differences between the distributions of measurement days with >10 hours of data were not detected between the reason and succinct groups, $\chi^2(7)=7.893$, p=0.342. A Bayesian contingency tables test provided $BF_{01}=6.96$ (Poisson sampling, prior concentration=1.0; prior robustness test depicts a concave function where, as concentration approaches 2, BF_{01} approaches 22.97).

Differences were not detected in valid wear day distributions between participants for whom reminders were sent, and for whom they were not: $\chi^2(7)=8.344$, p=0.303. A $BF_{01}=34.79$ (Poisson sampling, prior concentration=1.0; robustness function is concave as before. As concentration approaches 2, BF_{01} approaches 93.50).

Again, equivalence tests of mean differences between message types (-0.07 days, 90% CI [-0.47; 0.33]) was statistically significantly larger than d=-0.3 and smaller than d=0.3. The mean difference between reminder and opt out groups (-0.18 days, 90% CI [-0.60; 0.24]) was statistically significantly smaller than d=0.3, but we could not reject the hypothesis that the effect was higher than d=-0.3.

A MANOVA with both total wear time minutes and valid wear days as dependent variables neither detected differences between the reason, succinct and opt out groups ($F(4, 682)=2.335$, p=0.054, Wilk's $\Lambda=0.973$), although multicollinearity may have posed a problem to the model ($\tau=0.81$, $\rho=0.93$).

Dose dependence

If reading of messages is linearly related to wear time, an upward moving slope in means would have been expected. The dose dependence curve (Figure 7) is flat, showing no support for such a relationship between messages and wear time.

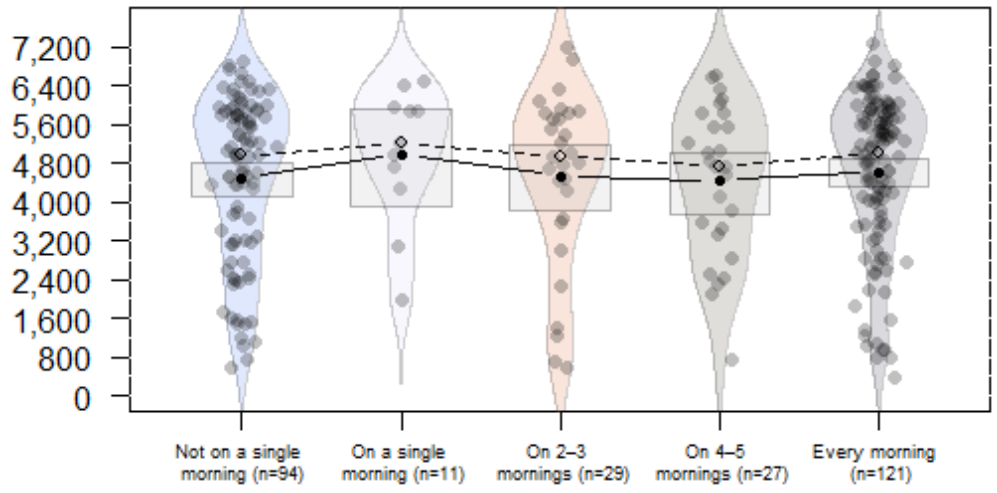


Figure 7: Self-reported opening and reading of messages. Y-axis is total wear time. Boxes represent 95% HDIs for the means, solid lines connect means and dashed lines connect medians. Participants who opted out of reminders are aggregated with those who indicated not having opened the messages even once. Participants who received messages, but did not answer the question on message reading, are excluded.

Discussion

This study evaluated the effects of two interventions to increase accelerometer wear times in the Let's Move It trial internal pilot study, and specifically, tested the effects of the because-heuristic on accelerometer wear time in older adolescents. We did not detect increased wear times among participants who received a reason in their daily SMS reminders, nor did we detect different wear times between those receiving the reminder messages and those who opted out. In all cases, null models were supported over those with small-to-medium sized effects.

Our main results are in line with the results of Belton et al. [11], of reminders not being able to increase wear time. This, although we attempted to improve on the earlier studies for example by not having exactly the same message sent every day. We do not have data on whether the reminder caused our participants put on the accelerometer more often, in spite of not increasing wear time [as found in 11].

Although the xerox machine study [12] has been highly publicised for thirty years now, the contextual framework of the effect remains unclear to an extent. Thus, the pool of possible reasons for null results is vast. These reasons would include e.g. the impersonal nature of SMS communication (as compared to face-to-face interaction), the source of the information, being incapable to complete the requested task, and a several other factors varying in plausibility (ranging from demographic factors to the presence of copy machines).

Bayesian analysis allowed us to quantify evidence for the null effect. The BF_{01} of 12.05 from testing equality of weartime means between the three arms (i.e. the Bayesian ANOVA) is enough to move an impartial observer with 1:1 prior odds to a

7.7% subjective posterior probability of an effect. Equally, a person who had high prior confidence (10 to 1 prior odds, translating to 91% probability) in the arms showing differences in weartimes, should become impartial and be moved to 54.6% probability in favor of an effect – provided that the proponent would agree with our methodology to test the hypothesis. (See supplement section “Interpreting Bayes Factors” for more information.)

Some might be tempted to interpret a "trend towards significance" from the MANOVA result, p=0.054. This would be an unjustified conclusion, as p-values are random variables and the only question is whether or not we are studying real phenomena or not – if we are, p-values near zero are always more likely than those near 0.05 [57]. Even if the arbitrary threshold of 0.05 would have been reached, it would have still been weak evidence; at best p=0.05 can indicate, that the H1 is two and a half times more likely than H0 (BF_{10}=2.45 in the *optimal* case; see [58], but also [59]).

Accordingly, the effect of reasons on this particular behavior, given our context and delivery method, has proved smaller than what would be considered minimally interesting (although participants did attribute remembering to wear their accelerometer to receipt of the reminders), and possibly zero. Thus, the *naïve because-heuristic* does not receive support in the current study. We can not make conclusions regarding the weaker claim of reasons affecting only tasks which are easy to carry out, due to design and sample size considerations.

The flat dose-dependence curve can indicate several things, including the possibility that text messages do not affect wear times. Attributing remembering to getting reminded could be a case of a *post hoc* reasoning error [60]. Another possibility is

that the messages could have had a small effect, but opening and reading the message provided no additional benefit. For example, the participant could have looked at the preview of the message on the cell phone screen and remembered without reading the whole message.

As there were no differences between the SMS and no-SMS arms, this effect may have been masked by selection bias, with those people who expect to experience problems with remembering, opting in to receive SMS reminders. As consent was almost fully dependent on the recruitment prompt, an additional assumption is needed that the two recruitment waves differ qualitatively (on an unobserved confounder). So, for example, the second wave may have consisted of more compliant participants or the potential interactions with the first wave participants might have made the opinion of the study more favorable. Thirdly, the effect of reminders may not have been linear, or only a small dose is needed to form a habit, and thus achieve maximal effect. This explanation requires the same assumptions as the one described above. Fourth, the flat curve may also be caused by unreliable measurement: dose should be operationalized in a way not dependent on self-report. Finally, it is possible that receiving reminders causes an undermining of one's own responsibility, so that those who receive reminders relinquish control and do not carry out the remembering techniques (e.g. placing of the accelerometer in a conspicuous place as a prompt to put it on) they would have, in the absence of reminders.

It may be that daily accelerometer wear is not determined by heuristic/automatic processes, but rather, is under more reflective reasoning processes. In this case, these reminders should have provided justifications and rationales that truly are important

for this target group. We do not have any evidence what thoughts and connotations our reminder content evoked in the youth's minds, and whether it was counterproductive. Finally, it is possible that participants who had agreed to take part in the accelerometer data collection already had made the reflective decision and proceeded to "implemental mindset" where persuasion messages are less relevant; e.g. as speculated in [61].

Limitations and strengths

There are a number of ways this study could have been improved on.

Opening and reading the messages (manipulation success)

Number of participants who opened and read the messages was assessed with a questionnaire instead of objective log data. This self-report measure (as well as the other post-intervention questionnaire items) was only a non-validated single item, thus probably far from optimal in terms of reliability. We had no reliable way to certify at which times the messages were received or whether they were opened at all. Anecdotal evidence indicated that the messages were too late for some students (i.e. they had already left the house and forgotten the accelerometer when receiving the message). On the other hand, we deemed sending the messages too early might pose an acceptability issue. The SMS queue in the gateway device presented a difficulty: larger number of message recipients heavily affected the deviation of delivery times, making the last messages in the queue arrive late for some students. During the second recruitment wave, time of initiating the send process was changed to be 45 minutes earlier (06:15 instead of 07:00), but we do not have data on the effect of this change.

We attempted to alleviate effects of not opening the messages by starting the each with the word “because”, so that message preview would render it visible on many devices even when not opened. Unfortunately we did not have access to a gateway system that could have sent e.g. MMS-messages, where a small picture could have been added, thus providing log data on how many times the picture was downloaded.

Contamination effects and masking the different message conditions

Participants may have found out their group allocation when discussing the messages with peers. This would require the discussion to have been about the nuances of message content and assumes that the participants are intrigued enough to spend time on making such inferences in the first place; an assumption perhaps not warranted. It is unclear how the discovery of SMS arm would have affected the results, but the possibility of confounding cannot be excluded. Randomising the groups by clusters could have helped to avoid this, but would have led to a reduction in statistical power. Still, the participants reported mainly not having discussed the messages with peers.

Sampling plan

The stopping rule for data collection was not defined in advance. The decision to collect another wave of participants with the same design was made, when it became apparent that the messages did not have the strong impact we had anticipated. This leads to uninformative p-values in terms of error control [62], whereas Bayesian analyses are not as crucially affected by stopping rules [63, but see also 64].

Lack of a randomised no-SMS control group

In order to avoid distortion of main trial outcomes (e.g. increased PA), care had to be taken in this within-trial RCT. The risk of sabotage due to disappointment of being allocated to a no-SMS control group was deemed too high, and thus participants were not randomised into a no-SMS group. This, in turn, lessens the strength of conclusions based on wear times between the participants receiving the reminder and those not receiving one. People who know they do not need a reminder may have thus ended up self-selecting to the no-SMS group.

This presumes that teenagers studying in a vocational school have the capacity to make accurate predictions about their future self-regulation capabilities in an unfamiliar task (putting on an accelerometer). On the other hand, as described, the wording of the recruitment prompt was slightly modified from wave 1 to wave 2, and consent to reminders was increased from 53% (85 out of 97) to 95% (176 out of 186), whereas wear times did not differ. Thus, strong selection effects seem unlikely. Although this indicates that opting out was more a result of the recruitment procedure than knowledge of not needing the reminders, future research should aim to randomise when feasible.

One way to address this problem would have been an n-of-1 design, where each day is randomised to one of the three message conditions. With this design, one should be careful to not leave learning effects undetected, as participants could habituate to reminders and forget in the concurrent absence of them.

Message content and size of request

The intervention was not piloted, nor was extensive testing of it's component parts done, which may have affected the results. The pre-testing of the message content was limited, too, and we thus do not have data on whether our participants considered the messages persuasive. This could be important theoretically, especially if the request size was considered large and our reasons were perceived as placebic or near-placebic. However, this might not be an issue in the first place, as participants had already agreed to wear the accelerometer as part of the trial.

Pre-registration

In this paper, we attempted to answer to the call of more stringent methodology by pre-registration. Optimally, this would have been done prior to beginning data collection. In these cases, it has been proposed that analyses should be considered exploratory [65]—especially in the presence of researcher degrees of freedom or data-dependent analysis decisions [66]—and can render p-values meaningless. In our case, this mistake turned out to be nonconsequential. We used Bayes factors to avoid claiming findings based on p-values alone, as recently warned against by the American Statistical Association [67]. Other approaches we used to address the replicability problem were transparent reporting and open data.

Rational theory defense

We must be careful not read too much into potential explanations (such as the hidden moderators-argument) for why an effect was not detected here. In the light of the recent "crisis of confidence in the psychological sciences" [68], it is concerning that

only a single direct replication of the xerox machine study has been published. The lack of direct replication and the mixed results from conceptual replications point to a more specific question in the context of current research: when is it rational to defend a theory by coming up with additional auxiliary hypotheses or rejecting the protocol of a falsifying experiment (falsification and corroboration being continuous measures, defined by the strictness of the test). Meehl [69] argues, from a neo-Popperian framework, for the Lakatos principle: *it is rational to defend a (seasoned) theory when it has accumulated an impressive track record of strong successes*.

As measured by Bayes Factors, even without accounting for possible publication bias, the LBC study does not reach the criterion for strong evidence (see data at https://osf.io/7y25w/). It would thus be quite a leap to consider the LBC theory (much less the stronger formulation by Cialdini and others) having accumulated enough credit by strong successes to justify much speculation about e.g. moderating factors.

Implications for practice

Our results, in line with some other studies [e.g. 11] indicate that researchers should not expect simple reminders to have strong effects on accelerometer wear times among youth. Also, despite previous strong claims, the because-heuristic in this context lacks the strength attributed to it in the popular literature. When considering using SMS reminders for youth, we suggest ensuring that remembering plausibly plays the key role in compliance with the behavior and target group in question, instead of other determinants/factors (such as social norms or motivation). Participants' coping skills and attention span may act as a ceiling to the potential effect of the reminder in situations where the target behavior can not immediately be

carried out, so suitability of SMS reminders could be assessed in these respects as well.

Implications for future research

To an extent, the findings here apply to situations where cost-effective reminders can potentially improve compliance. These areas may range from medication adherence [70] to sunscreen use [71]. An interesting hypothesis to test, would be whether reminders actually *reduce* active coping strategies that people use spontaneously – this could partly explain some null findings in the literature on technical reminder systems [72]. Second, the delivery of the reminders should optimally be objectively trackable, in order to make firm conclusions about the independent effects of delivery and receipt. Third, the context (including timing and location) where the participant receives the reminder is likely to be important, as well as the coping behaviour of the control group. It may also be worthwhile to gauge whether altering frequency of reminders affects the target behavior [70], or if the system can be made such that it adapts to the users and their environments [73].

Conclusion

In this research, we have found evidence against the assumed superiority of the naïve because-heuristic; providing reasons in simple compliance requests having a general persuasive effect on behaviour. By using Bayesian methods and equivalence testing, we were able to claim evidence of no effect for the because heuristic in this setting. Likewise, sending SMS reminders was not associated with improved accelerometer wear times. Although we did not randomise the no-SMS group, the changed

recruitment procedure plausibly accounts for majority of the selection effect, and a more potent explanation for the lack of differing weartimes is reaching the ceiling of the participants' ability to wear in the absence of very high motivation.

Our design had several limitations, which should be improved upon in future research. All in all, we remain pessimistic of the efficacy of the naïve because-heuristic and of simple reminders, even if they have a potent effect in participants' perceptions.

We conclude that despite strong claims, there is reason to consider the study of the because-heuristic a degenerating research programme [74], although there may be some contexts where the technique works as intended. Seeking to increase accelerometry wear time in participants may benefit from a design using the intervention mapping approach [75], including a plausible theoretical framework.

List of abbreviations

LMI: Let's Move It intervention to increase physical activity and decrease sedentary behavior in older adolescents

RCT: Randomised controlled trial

BF: Bayes Factor

SMS: Short Message Service, also known as "text messaging"

Declarations

Ethics approval and consent to participate

This study was approved by the Hospital District of Helsinki and Uusimaa, The Ethics Committee for gynaecology and obstetrics, pediatrics and psychiatry (decision number 367/13/03/03/2014). All participants consented to participate in the study.

Consent for publication

The manuscript does not contain any individual person's data.

Availability of data and material

Data and materials will be available at https://osf.io/tbyaz/, once the anonymisation process of the main trial's data is complete.

Competing interests

The authors declare that they have no competing interests.

Funding

The Let's Move It study, within which this study was conducted, was funded by the Ministry of Education and Culture, funding number 34/626/2012 (years 2012–14), and OKM/81/626/2014, (years 2015–17), as well as the Ministry of Social Affairs and Health, funding number 201310238 (years 2013–15). Finalisation of the manuscript was done under funding by the Academy of Finland (MH: grant number 295765, NH: grant number 285283). The funding bodies played no role in the design

of the study or writing the manuscript, nor the data collection, analysis, or interpretation.

Authors' contributions

Detailed authors' contributions are presented in the CRediT contributor role taxonomy (Appendix 1).

Acknowledgements

We would like to thank the research participants and the schools, as well as the research staff who aided in collecting the data.

References

1. Cain KL, Sallis JF, Conway TL, Van Dyck D, Calhoon L. Using accelerometers in youth physical activity studies: a review of methods. J Phys Act Health. 2013;10:437–450.

2. Matthews CE, Hagströmer M, Pober DM, Bowles HR. Best practices for using physical activity monitors in population-based research. Med Sci Sports Exerc. 2012;44:S68.

3. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act. 2008;5:56.

4. Zhuang J, Chen P, Wang C, Huang L, Zhu Z, Zhang W, et al. Characteristics of missing physical activity data in children and youth. Res Q Exerc Sport. 2013;84:S41–7.

5. Ward DS, Evenson KR, Vaughn A, Rodgers AB, Troiano RP. Accelerometer use in physical activity: best practices and research recommendations. Med Sci Sports Exerc. 2005;37:S582-8.

6. Audrey S, Bell S, Hughes R, Campbell R. Adolescent perspectives on wearing accelerometers to measure physical activity in population-based trials. Eur J Public Health. 2012;cks081.

7. Sirard JR, Slater ME. Compliance with wearing physical activity accelerometers in high school students. J Phys Act Health. 2009;6:S148.

8. Sallis JF, Saelens BE, Frank LD, Conway TL, Slymen DJ, Cain KL, et al. Neighborhood built environment and income: examining multiple health outcomes. Soc Sci Med. 2009;68:1285–93.

9. Barak S, Wu SS, Dai Y, Duncan PW, Behrman AL. Adherence to Accelerometry Measurement of Community Ambulation Poststroke. Phys Ther. 2014;94:101–10.

10. Toftager M, Kristensen PL, Oliver M, Duncan S, Christiansen LB, Boyle E, et al. Accelerometer data reduction in adolescents: effects on sample retention and bias. Int J Behav Nutr Phys Act. 2013;10:140.

11. Belton S, O'Brien W, Wickel EE, Issartel J. Patterns of non-compliance in adolescent field based accelerometer research. J Phys Act Health. 2013;10:1181–5.

12. Langer EJ, Blank A, Chanowitz B. The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction. J Pers Soc Psychol. 1978;36:635.

13. Pratkanis AR. Social influence analysis: An index of tactics. In: Pratkanis AR, editor. Sci Soc Influ Adv Future Prog. New York: Psychology Press; 2007. p. 17–82.

14. Cialdini RB, Goldstein NJ, Martin SJ. Influence: Science and practice. Boston: Pearson Education; 2009.

15. Blount J. Fanatical Prospecting: The Ultimate Guide to Opening Sales Conversations and Filling the Pipeline by Leveraging Social Selling, Telephone, Email, Text, and Cold Calling. John Wiley & Sons; 2015.

16. Goldman B. The Science of Settlement: Ideas for Negotiators. Pennsylvania: ALI-ABA; 2008.

17. Mortensen KW. Maximum Influence: The 12 Universal Laws of Power Persuasion. 2nd ed. New York: American Management Association; 2013.

18. Weinschenk S. The Power of the Word "Because" To Get People To Do Stuff [Internet]. Psychol. Today. 2013 [cited 2015 Nov 5]. Available from: https://web.archive.org/web/20170306230957/https://www.psychologytoday.com/blog/brain-wise/201310/the-power-the-word-because-get-people-do-stuff

19. Cialdini RB. Influence: Science and practice. 4th ed. USA: Arizona State University: Allyn & Bacon; 2001.

20. Key SM, Edlund JE, Sagarin BJ, Bizer GY. Individual differences in susceptibility to mindlessness. Personal Individ Differ. 2009;46:261–4.

21. Folkes VS. Mindlessness or mindfulness: A partial replication and extension of Langer, Blank, and Chanowitz. J Pers Soc Psychol. 1985;48:600–4.

22. Langer EJ, Chanowitz B, Blank A. Mindlessness–mindfulness in perspective: A reply to Valerie Folkes. J Pers Soc Psychol. 1985;48:605–7.

23. Makel MC, Plucker JA, Hegarty B. Replications in Psychology Research How Often Do They Really Occur? Perspect Psychol Sci. 2012;7:537–42.

24. Pollock CL, Smith SD, Knowles ES, Bruce HJ. Mindfullness Limits Compliance With the That's-Not-All Technique. Pers Soc Psychol Bull. 1998;24:1153–1157.

25. Slugoski BR. Mindless processing of requests? Don't ask twice. Br J Soc Psychol. 1995;34:335–350.

26. Laaksonen M, Talala K, Martelin T, Rahkonen O, Roos E, Helakorpi S, et al. Health behaviours as explanations for educational level differences in cardiovascular and all-cause mortality: a follow-up of 60 000 men and women over 23 years. Eur J Public Health. 2008;18:38–43.

27. Elgar FJ, Pförtner T-K, Moor I, De Clercq B, Stevens GWJM, Currie C. Socioeconomic inequalities in adolescent health 2002–2010: a time-series analysis of 34 countries participating in the Health Behaviour in School-aged Children study. The Lancet. 2015;385:2088–95.

28. National institute for Health and Welfare. School health survey 2015 results: Lifestyle [Internet]. Terveyden Ja Hyvinvoinnin Laitos. 2015 [cited 2015 Dec 4]. Available from: https://web.archive.org/web/20170306230805/https://www.thl.fi/fi/tutkimus-ja-asiantuntijatyo/vaestotutkimukset/kouluterveyskysely/tulokset/tulokset-aiheittain/elintavat

29. Hankonen N, Heino MTJ, Araujo-Soares V, Sniehotta FF, Sund R, Vasankari T, et al. 'Let's Move It' – a school-based multilevel intervention to increase physical activity and reduce sedentary behaviour among older adolescents in vocational secondary schools: a study protocol for a cluster-randomised trial. BMC Public Health. 2016;16:451–66.

30. Hankonen N, Heino MTJ, Hynynen S-T, Laine H, Araújo-Soares V, Sniehotta FF, et al. Randomised controlled feasibility study of a school-based multi-level intervention to increase physical activity and decrease sedentary behaviour among vocational school students. Int J Behav Nutr Phys Act [Internet]. 2017 [cited 2017 Mar 22];14. Available from: http://ijbnpa.biomedcentral.com/articles/10.1186/s12966-017-0484-0

31. Heino MTJ. No use reasoning with adolescents? A randomised controlled trial comparing persuasive messages [Internet]. 2016 [cited 2017 Jun 7]. Available from: https://helda.helsinki.fi/handle/10138/163800

32. Heino MTJ. Comparing persuasive SMS reminders: Supplementary website [Internet]. 2018 [cited 2018 Feb 21]. Available from: https://web.archive.org/web/20180221165832/https://heinonmatti.github.io/sms-persuasion/sms-persuasion-supplement.html

33. Dienes Z. Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. Palgrave Macmillan; 2008.

34. Wagenmakers E-J. A practical solution to the pervasive problems of p values. Psychon Bull Rev. 2007;14:779–804.

35. Fanelli D. Only Reporting Guidelines Can Save (Soft) Science. Eur J Personal. 2013;27:120–44.

36. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Extending the CONSORT Statement to Randomized Trials of Nonpharmacologic Treatment: Explanation and Elaboration. Ann Intern Med. 2008;148:295–309.

37. Allen L, Scott J, Brand A, Hlava M, Altman M. Publishing: Credit where credit is due. Nature. 2014;508:312–3.

38. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.

39. RStudio Team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA: RStudio, Inc.; 2015. Available from: http://www.rstudio.com/

40. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2009. Available from: http://ggplot2.org

41. Phillips N. yarrr: A companion to the e-book YaRrr!: The Pirate's Guide to R [Internet]. 2016. Available from: http://www.r-bloggers.com/the-new-and-improved-pirateplot-now-with-themes/

42. Baguley T. Serious stats: A guide to advanced statistics for the behavioral sciences. China: Palgrave Macmillan; 2012.

43. Bowman AW, Azzalini A. R package sm: nonparametric smoothing methods (version 2.2-5.4) [Internet]. University of Glasgow, UK and Università di Padova, Italia; 2014. Available from: URL http://www.stats.gla.ac.uk/ adrian/sm, http://azzalini.stat.unipd.it/Book_sm

44. Peters G-Jo. userfriendlyscience: Quantitative analysis made accessible [Internet]. 2016. Available from: http://CRAN.R-project.org/package=userfriendlyscience

45. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev [Internet]. 2015 [cited 2015 Oct 11]; Available from: http://link.springer.com/10.3758/s13423-015-0947-8

46. Heino MTJ, Vuorre M, Hankonen N. Bayesian evaluation of behavior change interventions: A brief introduction and a practical example. PsyArXiv [Internet]. 2017 [cited 2017 Dec 17]; Available from: https://psyarxiv.com/xmgwv/

47. Morey RD, Romeijn J-W, Rouder JN. The philosophy of Bayes factors and the quantification of statistical evidence. J Math Psychol [Internet]. 2016 [cited 2016 Jan

19]; Available from: http://www.sciencedirect.com/science/article/pii/S0022249615000723

48. Etz A, Vandekerckhove J. Introduction to Bayesian Inference for Psychology [Internet]. 2017 [cited 2017 Mar 21]. Available from: https://osf.io/preprints/psyarxiv/q46q3

49. Etz A, Vandekerckhove J. A Bayesian Perspective on the Reproducibility Project: Psychology. PLOS ONE. 2016;11:e0149794.

50. Morey RD, Rouder JN. BayesFactor: Computation of Bayes Factors for Common Designs [Internet]. 2015. Available from: https://CRAN.R-project.org/package=BayesFactor

51. Richard FD, Bond CF, Stokes-Zoota JJ. One Hundred Years of Social Psychology Quantitatively Described. Rev Gen Psychol. 2003;7:331–63.

52. Jamil T, Ly A, Morey RD, Love J, Marsman M, Wagenmakers E-J. Default "Gunel and Dickey" Bayes factors for contingency tables. Behav Res Methods. 2015;1–15.

53. Lakens D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Soc Psychol Personal Sci. 2017;8:355–62.

54. Lakens D. TOSTER: Two One-Sided Tests (TOST) Equivalence Testing [Internet]. 2016. Available from: https://CRAN.R-project.org/package=TOSTER

55. Gelman A, Carlin J. Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci. 2014;9:641–51.

56. Davis-Stober CP, Dana J. Comparing the accuracy of experimental estimates to guessing: a new perspective on replication and the "Crisis of Confidence" in psychology. Behav Res Methods. 2013;46:1–14.

57. Murdoch DJ, Tsai Y-L, Adcock J. P-Values are Random Variables. Am Stat. 2008;62:242–5.

58. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence. J Am Stat Assoc. 1987;82:112–122.

59. Mayo D, Morey RD. A Poor Prognosis for the Diagnostic Screening Critique of Statistical Tests. Open Sci Framew [Internet]. 2017 [cited 2018 Feb 16]; Available from: https://osf.io/ps38b/

60. Hansen H. Fallacies. In: Zalta EN, editor. Stanf Encycl Philos [Internet]. Summer 2015. 2015 [cited 2016 Mar 12]. Available from: https://plato.stanford.edu/entries/fallacies/

61. Armor DA, Taylor SE. The Effects of Mindset on Behavior: Self-Regulation in Deliberative and Implemental Frames of Mind. Pers Soc Psychol Bull. 2003;29:86–95.

62. Sagarin BJ, Ambler JK, Lee EM. An Ethical Approach to Peeking at Data. Perspect Psychol Sci. 2014;9:293–304.

63. Dienes Z. Using Bayes to get the most out of non-significant results. Quant Psychol Meas. 2014;5:781.

64. Simonsohn U. Posterior-Hacking: Selective Reporting Invalidates Bayesian Results Also [Internet]. Rochester, NY: Social Science Research Network; 2014 Jan. Report No.: ID 2374040. Available from: https://papers.ssrn.com/abstract=2374040

65. Wagenmakers E-J, Wetzels R, Borsboom D, Maas HLJ van der, Kievit RA. An Agenda for Purely Confirmatory Research. Perspect Psychol Sci. 2012;7:632–8.

66. Gelman A, Loken E. The Statistical Crisis in Science. Am Sci. 2014;102:460–5.

67. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;00–00.

68. Earp BD, Trafimow D. Replication, falsification, and the crisis of confidence in social psychology. Quant Psychol Meas. 2015;6:621.

69. Meehl PE. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. Psychol Inq. 1990;1:108–141.

70. Pop-Eleches C, Thirumurthy H, Habyarimana JP, Zivin JG, Goldstein MP, Walque DD, et al. Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders. AIDS Lond Engl. 2011;25:825.

71. Armstrong AW, Watson AJ, Makredes M, Frangos JE, Kimball AB, Kvedar JC. Text-message reminders to improve sunscreen use: a randomized, controlled trial using electronic monitoring. Arch Dermatol. 2009;145:1230–6.

72. Demonceau J, Ruppar T, Kristanto P, Hughes DA, Fargher E, Kardas P, et al. Identification and assessment of adherence-enhancing interventions in studies assessing medication adherence through electronically compiled drug dosing histories: a systematic literature review and meta-analysis. Drugs. 2013;73:545–62.

73. Hekler EB, Klasnja P, Riley WT, Buman MP, Huberty J, Rivera DE, et al. Agile science: creating useful products for behavior change in the real world. Transl Behav Med. 2016;6:317–28.

74. Lakatos I. History of science and its rational reconstructions [Internet]. Springer; 1971 [cited 2015 Dec 2]. Available from: http://link.springer.com/chapter/10.1007/978-94-010-3142-4_7

75. Eldredge LKB, Markham CM, Kok G, Ruiter RA, Parcel GS, others. Planning health promotion programs: an intervention mapping approach [Internet]. New Jersey: John Wiley & Sons; 2016 [cited 2016 Aug 31]. Available from: https://www.google.com/books?hl=en&lr=&id=UyrdCQAAQBAJ&oi=fnd&pg=PR11&dq=bartholomew+intervention+mapping&ots=Ocav5zMOvx&sig=M85aXJ_BAj0NXG79kwlK0fb5Ohs

Appendices

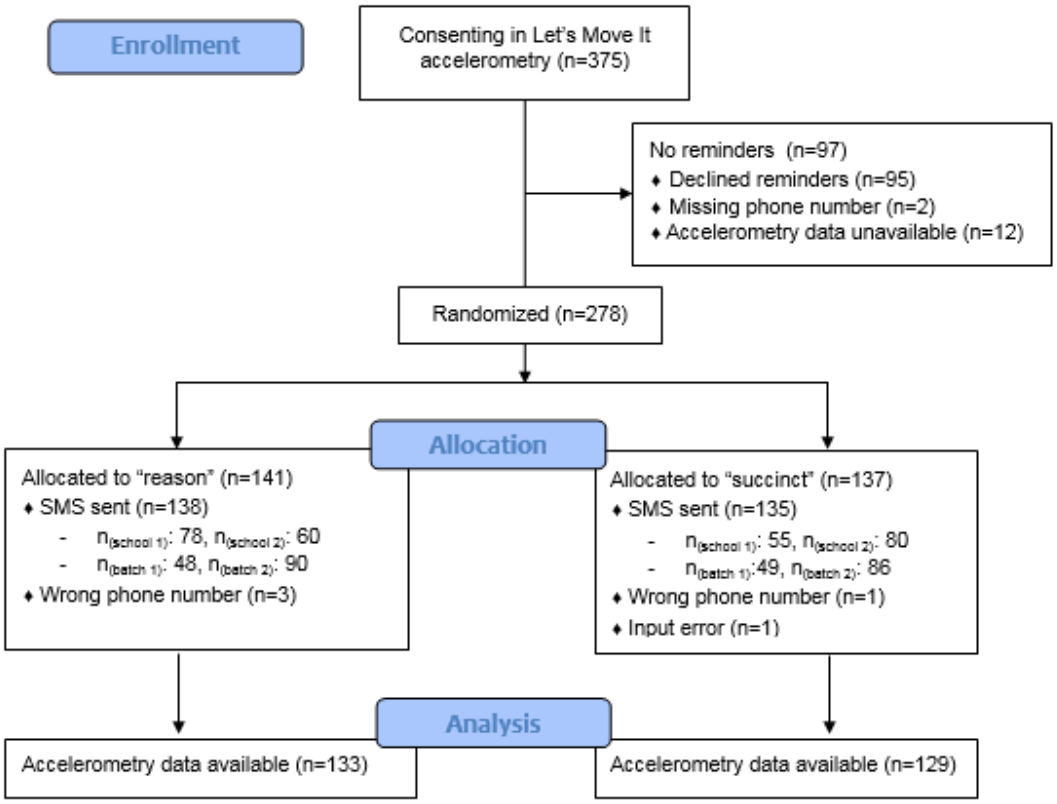
Appendix 1: CRediT – contributor role taxonomy

Taxonomy category	Description	Author responsible
Study conception	Ideas; formulation of research question; statement of hypothesis.	Nelli Hankonen & Matti Heino
Methodology	Development or design of methodology; creation of models.	Matti Heino, Nelli Hankonen, Keegan Knittle, Ari Haukkala
Computation	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms.	UKK-institute, Kryptoniitti joint-stock company
Formal analysis	Application of statistical, mathematical or other formal techniques to analyse study data.	Matti Heino
Investigation: performed the experiments	Conducting the research and investigation process, specifically performing the experiments.	Matti Heino & Let's Move It data collection team
Investigation: data/evidence collection	Conducting the research and investigation process, specifically data/evidence collection.	Let's Move It data collection team
Resources	Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation or other analysis tools.	Tommi Vasankari, Nelli Hankonen
Data curation	Management activities to annotate (produce metadata) and maintain research data for initial use and later re-use.	Matti Heino
Writing/manuscript preparation: writing the initial draft	Preparation, creation and/or presentation of the published work, specifically writing the initial draft.	Matti Heino
Writing/manuscript preparation: critical review, commentary or revision	Preparation, creation and/or presentation of the published work, specifically critical review, commentary or revision.	Matti Heino, Nelli Hankonen, Ari Haukkala, Keegan Knittle, Tommi Vasankari
Writing/manuscript preparation: visualization/data presentation	Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.	Matti Heino
Supervision	Responsibility for supervising research; project orchestration; principal investigator or other lead stakeholder.	Nelli Hankonen

Project administration	Coordination or management of research activities leading to this publication.	Nelli Hankonen
Funding acquisition	Acquisition of the financial support for the project leading to this publication.	Nelli Hankonen, Ari Haukkala, Tommi Vasankari, the UKK-institute

[37]

Appendix 2: CONSORT Flow Diagram



Appendix 3: Post-SMS questionnaire.

Nimi: _____

Ruksi sopiva:

Avasin tekstiviestin ja luin sen lähetysaamuna...

En yhtenäkään aamuna	Yhtenä aamuna	2-3 aamuna	4-5 aamuna	Jokaisena aamuna

Keskustelin tekstiviestien sisällöstä koulukaverieni kanssa...

En kertaakaan	Kerran	2-3 kertaa	4-5 kertaa	Useammin

Olin tyytyväinen viestien sisältöön:

Täysin eri mieltä	Jokseenkin eri mieltä	Ei samaa eikä eri mieltä	Jokseenkin samaa mieltä	Täysin samaa mieltä

Muuta palautetta tekstiviesteistä?

Iso kiitos! ☺