University of Groningen

Measuring steps with the Fitbit activity tracker

Dontje, Manon L; de Groot, Martijn; Lengton, Remko R; van der Schans, Cees P; Krijnen, Wim P

Published in:
Journal of medical engineering & technology

DOI:
[10.3109/03091902.2015.1050125](https://doi.org/10.3109/03091902.2015.1050125)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2015

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Dontje, M. L., de Groot, M., Lengton, R. R., van der Schans, C. P., & Krijnen, W. P. (2015). Measuring steps with the Fitbit activity tracker: an inter-device reliability study. *Journal of medical engineering & technology*, 39(5), 286-290. <https://doi.org/10.3109/03091902.2015.1050125>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Measuring steps with the Fitbit activity tracker: an inter-device reliability study

Manon L. Dontje, Martijn de Groot, Remko R. Lengton, Cees P. van der Schans & Wim P. Krijnen

To cite this article: Manon L. Dontje, Martijn de Groot, Remko R. Lengton, Cees P. van der Schans & Wim P. Krijnen (2015) Measuring steps with the Fitbit activity tracker: an inter-device reliability study, Journal of Medical Engineering & Technology, 39:5, 286-290, DOI: [10.3109/03091902.2015.1050125](https://doi.org/10.3109/03091902.2015.1050125)

To link to this article: <https://doi.org/10.3109/03091902.2015.1050125>

Published online: 27 May 2015.

Submit your article to this journal [↗](#)

Article views: 2088

View related articles [↗](#)

View Crossmark data [↗](#)

Citing articles: 16 View citing articles [↗](#)

INNOVATION

Measuring steps with the Fitbit activity tracker: an inter-device reliability study

Manon L. Dontje^{*1}, Martijn de Groot^{1,2}, Remko R. Lengton², Cees P. van der Schans^{2,3} and Wim P. Krijnen^{2,4}

¹Hanze University of Applied Sciences, Quantified Self Institute, Groningen, The Netherlands, ²Hanze University of Applied Sciences, Research and Innovation Group in Health Care and Nursing, Groningen, The Netherlands, ³Department of Rehabilitation, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, and ⁴Department of Statistics, University of Groningen, Groningen, The Netherlands

Abstract

Activity trackers like Fitbit are used for self-tracking of physical activity by an increasing number of individuals. Comparing physical activity scores with peers can contribute to the desired behavioural change. However, for meaningful social comparison a high inter-device reliability is paramount. This study aimed to determine the inter-device reliability of Fitbit activity trackers in measuring steps. Ten activity trackers (Fitbit Ultra) were worn by a single person (male, 46 years) during eight consecutive days. Inter-device reliability was assessed on three different levels of aggregation (minutes, hours, days) with various methods, including intra-class correlation coefficient (ICC), Bland-Altman plots, limits of agreement (LOA) and Mixed Model Analysis. Results showed that the inter-device reliability of the Fitbit in measuring steps is good at all levels of aggregation (minutes, hours, days), but especially when steps were measured per day. This implies that individuals can reliably compare their daily physical activity scores with peers.

Keywords

Accelerometry, Fitbit, objective measurement, physical activity, reliability

History

Received 13 February 2015

Revised 3 May 2015

Accepted 6 May 2015

1. Introduction

Regular physical activity has multiple beneficial health effects. For example, it reduces the risk of many chronic diseases such as cardiovascular disease and diabetes and it improves psychosocial well-being and quality-of-life [1,2]. Healthy adults are recommended to accumulate on average 10 000 steps per day to maintain or improve health [3]. Despite the well-known health benefits of regular physical activity, many people fail to meet the recommended level of physical activity [4].

To support the inactive population in increasing their physical activity, numerous activity trackers have been developed. Not only expensive accelerometers which can be used in research, but also the number of activity trackers accessible for individual consumers is increasing rapidly. Such activity trackers are often small, low-cost and user friendly. Activity trackers measure physical activity of the user, often expressed as the number of steps. Self-monitoring of daily physical activity may increase awareness about one's

own activity level, which is a pre-requisite for behavioural change [5,6].

By using activity trackers, people can compare their own physical activity with that of their peers, which can further enhance behavioural change. However, comparing scores requires sufficient inter-device reliability. Also, when activity trackers are being used in research, a high inter-device reliability is one of the pre-requisites [7]. Reliability of accelerometers is often examined by focusing on raw activity counts, but raw activity counts are often difficult to interpret, especially with regard to physical activity guidelines. Raw activity counts can be summarized and converted into more meaningful variables such as number of steps per minute, hour or day. For each of these derived variables it is important to determine the inter-device reliability, because each variable can be used for social comparison and research-related purposes.

One of the most popular recent activity trackers on the consumer market is Fitbit (Fitbit Inc., San Francisco, CA). To date, only little is known about the validity and reliability of this device [8,9]. Research suggests that the inter-device reliability for measuring steps on a treadmill is high [8,9], but there is no information on inter-device reliability for measuring steps in free-living conditions yet. Therefore, the aim of this study was to determine the inter-device reliability of Fitbit activity trackers in measuring steps during free-living conditions, over different units of time.

*Corresponding author. Email: m.l.dontje@alumnus.rug.nl

2. Methods

2.1. Design

Ten activity trackers (Fitbit Ultra, Fitbit Inc., San Francisco, CA) were worn by a single person (one of the researchers, male, 46 years) during 8 consecutive days, except during sleep and water-based activities (e.g. showering, bathing). The trackers were attached to the trouser, equally distributed over the left and right side pocket.

2.2. Measurements

The Fitbit (Fitbit Ultra) is a small device which uses a three dimensional accelerometer to measure movement of the user. The main outcome measure is steps, which can be combined with user data to calculate among others distance walked and energy expenditure. Scores are presented on the display and can be uploaded to the Fitbit website, where an overview of physical activity is presented together with a comparison with friends. The Fitbit API was used to extract all minute-to-minute data per day for each device. Since each device detects number of steps per minute, data per device consisted of the number of steps over $8 \times 24 \times 60 = 11\,520$ min in total.

2.3. Statistical analyses

Statistical analysis was performed by the statistical programming language R [10]; p values ≤ 0.05 were considered significant. Three levels of aggregation were distinguished; minutes, hours and days. For each aggregation level the reliability was examined in six steps, to be described subsequently below.

The first step was to examine whether the devices were equal in detecting movement vs no movement. Therefore, data were dichotomized (movement vs no movement) at the minute and hour level and the proportion of movement of each device was calculated. The proportions of movement of the 10 devices were compared with each other and equality was tested with the proportions test.

The second step was to obtain an impression on the degree of normality of the 10 activity trackers by analysing the measurements per minute by the Shapiro-Wilk normality test.

The third step was to calculate the two-way intra-class correlation coefficients (ICC) for absolute agreement in order to estimate the proportion of variance attributable to the devices [11]. The ICC's were calculated by comparing 45 pairs of distinct devices (10 devices make 45 possible twin combinations).

The fourth step was to determine the concordance correlation coefficient (CCC). This measure is comparable to the ICC, except that the CCC combines coefficients of precision and accuracy into a coefficient that measures the agreement to the identity line [12]. Values can range between -1 and 1 , indicating perfect disagreement to perfect agreement.

The fifth step was to visualise the agreement between each pair of Fitbits with Bland Altman plots. The limits of agreement (LOA) were also computed, to determine boundaries that contain 95% of the measurements [13,14]. The total deviation index (TDI) was computed to determine boundaries such that 95% of the measurements are within the boundary

from their target values [12]. Bland Altman plots were also used to visualise a possible increase of variability of the measurements with size, outliers, as well as existence of systematic bias.

The sixth step was to test any systematic effects in mean number of steps due to any of the devices. Therefore, Mixed Models Analyses were conducted, taking the devices as fixed effects and the aggregation level as random effects.

The afore-mentioned agreement coefficients were calculated for all 45 distinct pairs of Fitbits. There was no "gold standard" to compare each individual tracker with. Even so, to examine the reliability of the Fitbit we determined the agreement coefficients also by comparing the measurements of each Fitbit with the "true value" of measurement. The true value of measurement was approximated by the average of the remaining nine measurements. To obtain an impression of closeness between the average and each single Fitbit the mean difference (mean compared with the averaged remaining nine devices), percentage relative error ((mean difference/mean) * 100), ICC, CCC and its precision and accuracy measures, LOA and TDI were calculated. The LOA relative to the mean were calculated as well, to see whether there is an increase in precision with the aggregation level.

3. Results

The first step showed that the 10 devices were (almost) completely equal in detecting movement vs no movement when analysed in minutes as well as in hours ($p = 0.98$ and $p = 1.00$, respectively). This analysis was not applicable for aggregation level day, because there was no day without registered movement.

The Shapiro-Wilk normality tests indicated that normality of the measurements of the 10 activity trackers increased with level of aggregation. In 54% of the measurements of the 10 activity trackers over all minutes normality was not rejected, over all hours this was 84% and over all days this was 100%.

Third, the median (range) ICC was 0.90 (0.83–0.96), 1 (0.99–1) and 1 (0.99–1) for level of aggregation minutes, hours and days, respectively.

The CCC had a median (range) of 0.90 (0.83–0.96), 1 (0.99–1) and 0.99 (0.98–1) for level of aggregation minutes, hours and days, respectively.

Figure 1 presents examples of Bland Altman plots between two devices at three different levels of aggregation, including the corresponding LOA. For aggregation level minute, the median (range) LOA over the 45 distinct pairs of devices was 27.90 (17.47–36.86) steps, which indicates that 95% of the measurements were within the boundaries of 27.90 steps above and below the mean difference. For aggregation level hour, the median (range) LOA was 124.80 (62.62–229.80) steps. For aggregation level day, the median (range) LOA was 693.00 (228.40–1173.00) steps, which indicates that 95% of the measurements were within the boundaries of 693 steps above and below the mean difference. The median (range) TDI was 27.91 (17.54–36.89) steps, 129.00 (62.85–235.50) and 962.40 (244.6–1524.0) steps for level of aggregation minutes, hours and days, respectively.

The sixth step was to test any systematic effects in the number of steps depending on the different devices with

Figure 1. Bland Altman plots of two Fitbits at three different levels of aggregation: (A) Minutes; (B) Hours; (C) Days.

Mixed Model Analyses. There was a significant effect for devices when measured in minutes ($F=8.816$; $p<0.0001$), implying a significant systematic measurement difference. Five devices contributed significantly to the model, of which the size of the effect ranged from -1.33 to -0.52 steps (Table 1). There was also a significant effect for devices when measured in hours ($F=8.133$; $p<0.0001$). As shown in Table 1, four devices contributed significantly to the model, of which the size of the effect ranges from -16.26 to -32.78 steps. The Mixed Model Analysis also showed a significant effect for devices when measured in days ($F=4.328$; $p<0.001$). As shown in Table 1, only two devices contributed significantly to the model, of which the size of effect ranges from -499.88 to -295.63 steps.

The final step to examine reliability of the Fitbit was to calculate the agreement coefficients in comparison with

the “true value” of measurement. Per level of aggregation, the approximated true mean steps per minute, mean difference (mean compared with the averaged remaining nine devices), percentage relative error, LOA, TDI, CCC, ICC and the LOA relative to the mean are presented in Table 2. For aggregation level minutes, the LOA relative to the mean ranged between 0.44 – 0.79 , indicating that the LOA were $\sim$ times smaller in size than the mean number of steps per minute. For aggregation level hours, the LOA relative to the mean ranged between 0.07 – 0.24 , indicating that the LOA were more than 5-times smaller in size than the mean number of steps per hour. Also, for aggregation level days, the LOA relative to the mean ranged between 0.02 – 0.06 , indicating that the LOA were more than 25-times smaller in size than the mean number of steps per day.

Table 1. Estimation of systematic measurement error for individual Fitbit Ultra by mixed model analysis.

	SME steps/min	p Value	SME steps/hour	p Value	SME steps/day	p Value
(Intercept)	34.184	<0.001	840.877	<0.001	12 823.375	<0.001
Fitbit2	0.308	0.231	7.582	0.251	115.625	0.405
Fitbit3	-0.011	0.965	-0.279	0.966	-4.250	0.976
Fitbit4	0.406	0.115	9.984	0.130	152.250	0.274
Fitbit5	-0.788*	0.002	-19.385*	0.003	-295.625*	0.036
Fitbit6	-1.333*	0.000	-32.779*	0.000	-499.875*	0.001
Fitbit7	-0.677*	0.009	-16.664*	0.012	-254.125	0.070
Fitbit8	-0.406	0.115	-9.984	0.130	-152.250	0.274
Fitbit9	-0.517*	0.045	-12.721	0.054	-194.000	0.164
Fitbit10	-0.661*	0.010	-16.262*	0.014	-248.000	0.077

Fitbit1 was used as a reference for each of the three aggregation levels (minute, hour and day).

*Significant fixed effect of device.

Table 2. Reliability of single Fitbit Ultra's compared to the true value* at three levels of aggregation (minutes, hours and days).

	Fitbit1	Fitbit2	Fitbit3	Fitbit4	Fitbit5	Fitbit6	Fitbit7	Fitbit8	Fitbit9	Fitbit10
Steps/minute										
Mean	34.184	34.493	34.173	34.590	33.396	32.852	33.507	33.778	33.667	33.523
Mean difference	-0.409	-0.751	-0.396	-0.860	0.467	1.072	0.344	0.042	0.166	0.326
% Relative error	1.196	2.177	1.159	2.486	1.398	3.263	1.027	0.124	0.493	0.972
ICC	0.943	0.942	0.969	0.939	0.908	0.965	0.960	0.920	0.963	0.932
CCC	0.943	0.942	0.969	0.939	0.908	0.964	0.960	0.920	0.963	0.932
Precision	0.945	0.943	0.970	0.941	0.909	0.965	0.961	0.921	0.964	0.933
Accuracy	0.998	0.998	0.999	0.998	1.000	0.999	0.999	0.999	0.999	0.999
LOA	20.939	21.156	15.199	21.674	26.276	16.241	17.402	24.549	16.641	22.891
TDI	20.954	21.207	15.218	21.739	26.291	16.376	17.414	24.549	16.644	22.899
Relative LOA	0.613	0.613	0.445	0.627	0.787	0.494	0.519	0.727	0.494	0.683
Steps/hour										
Mean	840.877	848.459	840.598	850.861	821.492	808.098	824.213	830.893	828.156	824.615
Mean difference	-10.057	-18.481	-9.747	-21.149	11.483	26.364	8.459	1.036	4.078	8.013
% Relative error	1.196	2.178	1.160	2.489	1.398	3.262	1.026	0.125	0.492	0.972
ICC	0.998	0.999	0.999	0.995	0.999	0.999	1.000	0.999	0.999	0.999
CCC	0.998	0.999	0.999	0.995	0.999	0.999	1.000	0.999	0.999	0.999
Precision	0.998	0.999	0.999	0.996	0.999	0.999	1.000	0.999	0.999	0.999
Accuracy	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
LOA	142.801	95.210	67.122	199.993	77.931	85.772	58.294	82.913	79.619	96.183
TDI	144.163	101.919	69.808	204.275	81.140	100.244	60.624	82.936	80.021	97.466
Relative LOA	0.170	0.112	0.080	0.235	0.095	0.106	0.071	0.100	0.096	0.117
Steps/days										
Mean	12 823.375	12 939.000	12 819.125	12 975.625	12 527.750	12 323.500	12 569.250	12 671.125	12 629.375	12 575.375
Mean difference	-153.361	-281.833	-148.639	-322.528	175.111	402.056	129.000	15.806	62.194	122.194
% Relative error	1.196	2.178	1.160	2.486	1.398	3.262	1.026	0.125	0.492	0.972
ICC	0.995	0.998	0.999	0.995	0.999	0.998	1.000	0.998	0.998	0.996
CCC	0.995	0.996	0.998	0.992	0.998	0.993	0.999	0.998	0.998	0.996
Precision	0.998	0.999	0.999	0.995	0.999	0.998	1.000	0.999	0.998	0.996
Accuracy	0.997	0.997	0.999	0.996	0.999	0.994	0.999	1.000	1.000	1.000
LOA	799.085	519.595	446.023	838.073	357.948	514.452	188.781	475.538	539.055	700.348
TDI	861.261	786.566	543.990	1076.583	512.585	987.080	329.689	476.681	554.573	745.669
Relative LOA	0.062	0.040	0.035	0.065	0.029	0.042	0.015	0.038	0.043	0.056

*The true value of measurement was approximated by the average of the remaining nine measurements.

LOA, limits of agreement; TDI, total deviation index; CCC, concordance correlation coefficient; ICC, intra-class correlation coefficient; Relative LOA, limits of agreement relative to the mean.

4. Discussion

The aim of this study was to determine the inter-device reliability of Fitbit Ultra activity trackers in measuring steps during free-living conditions, over different units of time (minutes, hours, days). The results show that the inter-device reliability was (reasonably) good at all levels of aggregation, but especially when steps were counted per day. This indicates that social comparison is reliable for daily

measurements and that the Fitbit Ultra is replaceable by another within reliable standards.

This is based on rigorous analysis using various methods for assessing reliability of 10 Fitbit activity trackers. The reliability of the devices was examined at three different levels of aggregation over time, i.e. minutes, hours and days, with various methods. First, the results indicate that the 10 devices were equal in detecting movement vs no movement.

This means roughly that, when the user was moving, it was reliably detected by each device, whether activity was measured in steps per minute or per hour. The results showed that most reliability coefficients improved with increasing level of aggregation (from minutes to days). For example, the degree of normality increased from 54% (aggregation level minutes) to 100% (aggregation level days). The ICCs and CCCs increased from 0.90 to unity with increasing levels of aggregation. This indicates that the agreement between the different devices is excellent, especially for measuring steps per day. This is in line with other studies [8,9]. Generally, $ICC > 0.80$ and $CCC > 0.95$ are considered acceptable levels of agreement [15–17]. The increasing reliability with increasing level of aggregation could also be concluded from the Bland–Altman plots, as the number of outliers decreased when the level of aggregation increased from minute to day. Moreover, the LOA relative to the mean decreased enormously as the level of aggregation increased from minute to day. When steps were counted per minute, the LOA were 2-times smaller than the mean number of steps. When steps were counted per hour, the LOA were 5-times smaller than the mean number of steps. Also, when steps were counted per day, the LOA were ~25-times smaller than the mean number of steps.

The aforementioned methods showed reasonably high reliability, although the mixed model analyses indicated significant Fitbit effects for a minority of cases. All of these seem small in size from a clinical point of view. The maximum effects are ~1, 26 and 402 steps for aggregation level minute, hour and day, respectively. The maximum difference for all devices is 3.3%, which seems clinically acceptable. It comes close to the 3% error of miscounting which is maximally accepted by the quality standards of the Japanese industry of pedometers [18]. In general, an error percentage of less than 10% is considered acceptable for step counters [18]. Another study also found a low measurement error for Fitbits [9]. They compared steps of Fitbits to manual step count during walking on a treadmill and found that the measurement error was less than 1.3%. Taking the results of both studies together, the measurement error of Fitbit activity trackers seems acceptable.

When interpreting the results of the present study, it should be taken into account that only 10 activity trackers were examined. However, due to the design of this study 45 pairs of measurements could be examined with various methods. Nevertheless, for future research it is recommended to include a larger number of activity trackers, on a range of different subjects and activities for inter- and intra-device comparisons. A strength of the current study is that reliability was examined with several methods which all have their own distinct value, but when used together are more worthwhile than the sum of parts. Another strength is that the 8 days of measurement included different types of physical activity in real life, from light intensity physical activity to vigorous physical activity.

In conclusion, the inter-device reliability of the Fitbit Ultra in measuring steps in free-living conditions is good,

especially when measured per day. This implies that social comparison is reliable for daily measurements and that the Fitbit Ultra can be replaced by another Fitbit Ultra within reliable standards.

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

- Warburton, D.E., Nicol, C.W., and Bredin, S.S., 2006, Health benefits of physical activity: The evidence. *Canadian Medical Association Journal*, **174**, 801–809.
- Pedersen, B., and Saltin, B., 2008, Evidence for prescribing exercise as therapy in chronic disease. *Scandinavian Journal of Medicine & Science in Sports*, **16**, 3–63.
- Tudor-Locke, C., Hatano, Y., Pangrazi, R.P., and Kang, M., 2008, Revisiting ‘‘how many steps are enough?’’. *Medicine & Science in Sports & Exercise*, **40**, S537–S543.
- Tucker, J.M., Welk, G.J., and Beyler, N.K., 2011, Physical activity in US adults: Compliance with the physical activity guidelines for Americans. *American Journal of Preventive Medicine*, **40**, 454–461.
- Prochaska, J., and Norcross, J., 2001, Stages of change. *Psychotherapy*, **38**, 443–448.
- Weinstein, N.D., 1988, The precaution adoption process. *Health Psychology*, **7**, 355–386.
- Welk, G., Schaben, J., and Morrow, J., 2004, Reliability of accelerometry-based activity monitors: A generalizability study. *Medicine & Science in Sports & Exercise*, **36**, 1637–1645.
- Adam Noah, J., Spierer, D.K., Gu, J., and Bronner, S., 2013, Comparison of steps and energy expenditure assessment in adults of fitbit tracker and ultra to the actual and indirect calorimetry. *Journal of Medical Engineering & Technology*, **37**, 456–462.
- Takacs, J., Pollock, C.L., Guenther, J.R., Bahar, M., Napier, C., and Hunt, M.A., 2014, Validation of the fitbit one activity monitor device during treadmill walking. *Journal of Science & Medicine in Sport*, **17**, 495–500.
- R Development Core Team, 2011, *R: A Language and Environment for Statistical Computing*. [computer program] (Vienna, Austria: R Foundation for Statistical Computing).
- McGraw, K., and Wong, S., 1996, Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, **1**, 30–46.
- Lin, L., 2000, Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, **19**, 255–270.
- Altman, D.G., and Bland, J.M., 1983, Measurement in medicine: The analysis of method comparison studies. *The Statistician*, **32**, 307–317.
- Bland, M.J., and Altman, D., 1986, Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, **327**, 307–310.
- Lin, L.I., 1989, A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268.
- McBride, G., 2005, A proposal for strength-of-agreement criteria for lin's concordance correlation coefficient. Report nr NIWA Client Report: HAM2005-062.
- Shrout, P., and Fleiss, J., 1979, Intraclass correlations - uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420–428.
- Tudor-Locke, C., Sisson, S., Lee, S., Craig, C., Plotnikoff, R., and Bauman, A., 2006, Evaluation of quality of commercial pedometers. *Canadian Journal of Public Health-Revue Canadienne De Sante Publique*, **97**, S10–S15.