Original Paper

# Investigating Measurement Equivalence of Smartphone Sensor–Based Assessments: Remote, Digital, Bring-Your-Own-Device Study

Lito Kriara, PhD; Frank Dondelinger, PhD; Luca Capezzuto, PhD; Corrado Bernasconi, MD, PhD; Florian Lipsmeier, PhD; Adriano Galati, PhD; Michael Lindemann, PhD

F. Hoffmann-La Roche Ltd, Basel, Switzerland

**Corresponding Author:**
Lito Kriara, PhD
F. Hoffmann-La Roche Ltd
Grenzacherstrasse 124
Basel, CH-4070
Switzerland
Phone: 41 61 687 10 20
Email: lito.kriara@roche.com

## Abstract

**Background:** Floodlight Open is a global, open-access, fully remote, digital-only study designed to understand the drivers and barriers in deployment and persistence of use of a smartphone app for measuring functional impairment in a naturalistic setting and broad study population.

**Objective:** This study aims to assess measurement equivalence properties of the Floodlight Open app across operating system (OS) platforms, OS versions, and smartphone device models.

**Methods:** Floodlight Open enrolled adult participants with and without self-declared multiple sclerosis (MS). The study used the Floodlight Open app, a “bring-your-own-device” (BYOD) solution that remotely measured MS-related functional ability via smartphone sensor–based active tests. Measurement equivalence was assessed in all evaluable participants by comparing the performance on the 6 active tests (ie, tests requiring active input from the user) included in the app across OS platforms (iOS vs Android), OS versions (iOS versions 11-15 and separately Android versions 8-10; comparing each OS version with the other OS versions pooled together), and device models (comparing each device model with all remaining device models pooled together). The tests in scope were Information Processing Speed, Information Processing Speed Digit-Digit (measuring reaction speed), Pinching Test (PT), Static Balance Test, U-Turn Test, and 2-Minute Walk Test. Group differences were assessed by permutation test for the mean difference after adjusting for age, sex, and self-declared MS disease status.

**Results:** Overall, 1976 participants using 206 different device models were included in the analysis. Differences in test performance between subgroups were very small or small, with percent differences generally being ≤5% on the Information Processing Speed, Information Processing Speed Digit-Digit, U-Turn Test, and 2-Minute Walk Test; <20% on the PT; and <30% on the Static Balance Test. No statistically significant differences were observed between OS platforms other than on the PT (*P*<.001). Similarly, differences across iOS or Android versions were nonsignificant after correcting for multiple comparisons using false discovery rate correction (all adjusted *P*>.05). Comparing the different device models revealed a statistically significant difference only on the PT for 4 out of 17 models (adjusted *P*≤.001-.03).

**Conclusions:** Consistent with the hypothesis that smartphone sensor–based measurements obtained with different devices are equivalent, this study showed no evidence of a systematic lack of measurement equivalence across OS platforms, OS versions, and device models on 6 active tests included in the Floodlight Open app. These results are compatible with the use of smartphone-based tests in a bring-your-own-device setting, but more formal tests of equivalence would be needed.

***(J Med Internet Res 2025;27:e63090)*** doi: 10.2196/63090

**KEYWORDS**

Floodlight Open; multiple sclerosis; smartphone; sensors; mobile phone; wearable electronic devices; digital health; equivalence; device equivalence; cognition; gait; upper extremity function; hand motor function; balance; digital biomarker; variability; mHealth; mobile health; autoimmune disease; motor; digital assessment

## Introduction

Multiple sclerosis (MS) is a chronic, demyelinating autoimmune disease of the central nervous system, which can manifest in functional impairment in cognitive and motor abilities, and the subsequent accumulation of disability over time [1]. The number of people affected by MS is increasing globally, with recent figures estimating over 2.8 million cases worldwide [2]. While assessments of functional ability can inform medical decisions and interventions that can ultimately reduce the risk of relapses and slow down the rate of disease progression, their utility has been limited by their infrequent use and reliance on patient recall [3-5]. An objective and more frequent assessment could provide a more detailed picture of the evolution of the disease.

To address this unmet need, smartphone sensor–based tests, or remote digital assessment technologies, are increasingly being studied for assessing functional ability in people with MS [6-14]. Typically, such tests can be remotely and frequently performed in the patient's home environment without supervision by a health care professional [10,12,15,16]. Furthermore, the use of wearable or embedded sensors allows many different aspects of functional ability to be characterized and objectively quantified [12,13,17,18]. Thus, they provide more granular and more detailed information than captured with the single scores of traditional standard clinical assessments such as the Nine-Hole Peg Test, oral Symbol Digit Modalities Test, or the Timed 25-Foot Walk [19-21]. Ultimately, the goal of digital remote assessment tools in MS is to help uncover insidious disease progression [22-24] and allow timely and appropriate medical intervention, which could lead to better treatment outcomes.

While restricting remote digital assessment technologies to a specific device or a single operating system (OS) platform can offer a simpler integration of hardware and software [12,13,25], it may limit their uptake [26]. Bring-your-own-device (BYOD) solutions, by comparison, can help increase access to remote digital technologies (ie, reaching more users) by taking into account the differences in market share of iOS and Android devices across geographies [27]. BYOD solutions are also less affected by the unfamiliarity associated with novel tools, speak to the patients' preference of using their own device rather than using and carrying an additional device with them, and potentially improve the user experience [26,28,29]. Furthermore, they are associated with fewer logistical challenges and, as a result, are more cost-effective to deploy in clinical trials [30,31]. However, measurement equivalence across different device models must be first established before a BYOD solution can be successfully deployed [11,32].

The Floodlight Open app assesses MS-related functional ability through smartphone sensor–based "active" tests (ie, assessments that require active input from the user) and was specifically developed for use in a BYOD setting [33]. It was deployed in Floodlight Open, a global, open-access, digital-only study that was designed to understand the drivers and barriers in the deployment and persistence of use of a smartphone app in a naturalistic setting and broad study population [33]. Previously, it was shown that the Floodlight Open app can differentiate and discriminate between MS participants and non-MS participants [33,34]. Using data from this study, we sought here to establish the properties of measurement equivalence on 3 separate levels: OS platforms, OS versions, and device models.

## Methods

### Study Design and Participants

Floodlight Open was run in 17 countries, and data were collected between April 23, 2018, and April 26, 2023. The study design and the Floodlight Open app have been previously described [33]. As a fully remote, digital-only study, Floodlight Open did not involve supervision by health care professionals. Adult participants aged 18 years and older with or without self-declared MS living in one of the participating countries could enroll by providing full electronic consent and downloading the Floodlight Open app on their own smartphone device. After providing their electronic consent, participants received a token, or activation code, via email with which they could unlock the functionalities of the app. Participants were excluded from the analyses if they did not complete at least one valid active test or had missing device information.

### Ethical Considerations

The protocol, the electronic informed consent forms, data protection, and relevant supporting information were reviewed and approved by local institutional review boards or ethics committees before the study was initiated, as applicable, in accordance with each country's regulatory requirements. For example, the institutional review board for the United States was the Western Institutional Review Board in Puyallup, Washington (approval: 20180617). Further details on the institutional review boards and ethics committees' approvals are described in a previous report [33].

### Floodlight Smartphone Sensor-Based Active Tests

The Floodlight Open app was designed to remotely measure, in a naturalistic setting, functional ability in cognition, hand motor function, gait and balance, mobility, and mood through smartphone-based active tests, passively collected life-space measurements, and patient-reported outcomes. The individual tests have been previously described [33]. It supports all devices running iOS version 11.x or later or Android version 7.x or later, which were commonly available at the time the app was launched. The iOS version was first released on April 22, 2018, and the Android version on July 17, 2019.

The measurement equivalence properties were studied on 6 active tests, including the Information Processing Speed (IPS), Information Processing Speed Digit-Digit (IPS DD), Pinching Test (PT), Static Balance Test (SBT), U-Turn Test (UTT) and 2-Minute Walk Test (2MWT; Table 1). These active tests could be performed up to once daily (PT, SB, UTT, and 2MWT) or up to once weekly (IPS and IPS DD).

**Table 1.** Active tests included in the analysis.

| Functional domain and active test | Sensors used | Test schedule | Test feature to assess test performance | Quality control flags | |
|---|---|---|---|---|---|
| | | | | Test marked invalid if characterized by | Criterion |
| **Cognition** | | | | | |
| IPS[a] | Touchscreen | Weekly | Number of correct responses (n) | "Play-to-quit" attempt | • Response selected independently of symbol to be matched [15] |
| IPS DD test[b] | Touchscreen | Weekly | Number of correct responses (n) | "Play-to-quit" attempt | • Response selected independently of symbol to be matched |
| **Hand motor function** | | | | | |
| PT[c] | Touchscreen | Daily | Number of pinches (n) | "Play-to-quit" attempt | • No gestures recorded by the touch screen (no screen interaction) [15] |
| **Gait and balance** | | | | | |
| **SBT[d]** | | | | | |
| | Accelerometer | Daily | Sway path, $(m/s^2)$ | "Play-to-quit" attempt | • Phone is kept on the table [15]<br>• Steps recorded during the test [15] |
| **UTT[e]** | | | | | |
| | Accelerometer, Gyroscope | Daily | Turn speed, (rad/s) | "Play-to-quit" attempt | • Phone is kept on the table [15]<br>• Main orientation of the phone is stable for ≤90% of the time |
| | Accelerometer, Gyroscope | Daily | Turn speed, (rad/s) | Insufficient data | • Number of turns ≤ 3<br>• Turn angle is either ≥270 or ≤90 degrees<br>• Test duration is ≤40 s |
| **2MWT[f]** | | | | | |
| | Accelerometer | Daily | Number of steps (n) | "Play-to-quit" attempt | • Phone is kept on the table [15] |
| | Accelerometer | Daily | Number of steps (n) | Insufficient data | • Test duration ≤105 s |

[a] IPS: Information Processing Speed.

[b] IPS DD: Information Processing Speed Digit-Digit.

[c] PT: Pinching Test.

[d] SBT: Static Balance Test.

[e] UTT: U-Turn Test.

[f] 2MWT: 2-Minute Walk Test.

The IPS measured cognitive function and instructed participants to match as many symbols to digits as possible within 90 seconds according to a symbol-digit key provided at the top end of the smartphone display. To account for the visuomotor component involved in this substitution task, participants were additionally asked to match digits to digits instead during the 30-second IPS DD. The PT evaluated the ability to perform upper extremity function tasks. The goal was to pinch, or squeeze, as many tomato shapes as possible within 30 seconds. After each pinched tomato shape, a new shape appeared at a different location on the smartphone display. Gait and balance were assessed with 3 different active tests. The SBT instructed participants to stand as still as possible for 30 seconds with both feet on the ground and with their eyes open. The UTT prompted participants to perform at least 5 U-turns on an even ground 4 meters apart within 60 seconds and, thus, assesses both gait and dynamic balance. By comparison, the 2MWT assessed gait during straight walking without turning on an even ground for 2 minutes. For both the UTT and 2MWT, study participants were instructed to walk as fast as possible but safely; use of assistive devices or orthotics was permitted as needed and recorded.

The raw sensor data (touchscreen, accelerometer, and gyroscope data) were encrypted and transferred wirelessly to a secure central database server that is controlled and maintained by the study initiator, F. Hoffmann-La Roche Ltd. Subsequently, the raw sensor data were used to compute predefined test features that characterize the participant's performance on the individual active tests. These features include the number of correct responses on the IPS and IPS DD, the number of successful pinches on the PT [15], the sway path on the SBT [15], the average turn speed on the UTT [15], and the number of steps on the 2MWT.

## Data Processing and Statistical Analysis

Two data processing steps were implemented to reduce any potential bias introduced by poor sampling frequency of the embedded sensors or by nonaccordant test execution. The second data processing step, however, was not available for the Draw a Shape Test; consequently, this active test is not included in the analyses presented here.

In the first data processing step, any active test recorded with either a low sampling frequency (sampling frequency ≤33 Hz) or an unstable sampling frequency (sampling frequency temporarily dropped ≤33 Hz) was identified and subsequently excluded from the analyses since low or unstable sampling frequency can negatively affect sensor-based measurements [35]. The minimum sampling frequency was set at 33 Hz for all sensor types, as this is the lowest sampling frequency to reliably assess gait [36]. Most currently available commercial smartphone devices support this sampling frequency. Next, any active test that was not executed according to the test's instructions was disregarded as they were considered invalid. Such nonaccordant tests were retrospectively identified with quality control flags, which were derived from the raw sensor data and provided objective information on how the test was executed. Different quality control flags were defined for each of the active tests (Table 1). These quality control flags identified, without any input from an observer, individual tests that were either characterized by "play-to-quit" behavior or insufficient data. "Play-to-quit" behavior captures instances where the participant did not intend to perform the test according to the instructions provided and wanted to skip the test instead (eg, the same response is selected in fast succession irrespective of the symbol shown during the IPS or IPS DD; no touch screen interaction during the PT). The insufficient data criterion was introduced to ensure the extraction of meaningful gait features, which is particularly important when comparing smaller subgroups that are more sensitive to noise. According to this criterion, only gait tests with a sufficient number of steps or turns were kept for further analysis.

Measurement equivalence signifies that the measurements obtained from 2 groups stem from the same distribution, that is, the 2 groups are equivalent to each other. This was studied separately for each of the 6 active tests in both MS and non-MS participants using the Floodlight feature values, that is, measurements of test performance, derived from the raw sensor data (raw accelerometer, gyroscope, or touchscreen data). Since the study duration was not fixed, the first valid test execution (ie, the first valid IPS, IPS DD, PT, SBT, UTT, and 2MWT) of each participant was used to assess measurement equivalence in order to maximize the number of participants available for analysis and to compare like with like. A test run was considered completed if the participants performed all active tests except the 2MWT in a single session in a predefined, fixed sequence. As the 2MWT was self-administered independently from this sequence, the first valid 2MWT of each participant was used to study the equivalence of this active test.

Feature values derived from the raw sensor data (ie, raw accelerometer, gyroscope, or touchscreen data collected while the participants were performing the active tests) were adjusted for age, sex, and self-declared disease status ("MS" and "Non-MS") with a robust linear model to account for differences in these covariates (function rlm{} included in the Python *statsmodel* package version 0.13.5). Next, subgroups were defined in 4 separate categories: OS platform (ie, iOS and Android), iOS version (ie, iOS versions 11, 12, 13, 14, and 15), Android version (Android versions 8, 9, and 10), and device model. To minimize potential issues with small sample sizes but also allow for as many different subgroups (eg, device models) as possible to be included in the analysis, only subgroups with at least 20 participants were included.

To study measurement equivalence, each subgroup (eg, iOS 12) was subsequently compared against their respective reference group. This reference group consisted of all remaining subgroups of the same category pooled together. For example, the reference group for iOS 12 is iOS 11, 13, 14, and 15 pooled together. The null hypothesis was that the distribution of the subgroup—for example, iOS 12—is identical to the distribution of its respective reference group. Differences in the means between the subgroup under study and its reference group were tested for statistical significance through permutation testing [37,38]. This test was chosen as it is a nonparametric test that requires only minimal model assumptions and additionally is robust. It does not assume a normal distribution and is also robust with regard to unbalanced data sets such as ours and to outliers. To perform the permutation test, the overall population (in the example above, all iOS devices) was randomly sampled to generate 2 permuted groups whose sample sizes were identical to the subgroup under study (eg, iOS 12) and its reference group (eg, iOS 11, 13, 14, and 15 pooled together), respectively. This was repeated 10,000 times, resulting in 10,000 permutations, and hence 10,000 mean differences. The reported *P* value is defined as the proportion of permutations that resulted in a mean difference between the 2 permuted groups that is larger than the real observed mean difference between the subgroup and its reference group. Both unadjusted and adjusted *P* values (adjusted for multiple comparisons with false discovery rate correction using the Benjamini-Hochberg method) are reported.

Additional reported metrics include the 95% CIs of expected mean differences under the assumption of the null hypothesis, which were derived from the permutation test; the absolute difference and percent difference in the observed mean scores (mean feature values) between each subgroup and their respective reference group; the SD of each subgroup and its reference group; as well as the effect size of these differences (Cohen *d*; very small effect size: $d$=0.01; small effect size: $d$=0.2; medium effect size: $d$=0.5; large effect size: $d$=0.8) [39,40].

To corroborate the findings, a sensitivity analysis was run, which used the median as the test statistic difference (instead of the mean difference) of the permutation test.

In a separate analysis, smartphones were compared against tablets to evaluate the impact of screen size on measurement equivalence. Differences between smartphones and tablets were assessed for statistical significance with the Mann-Whitney *U* Test.

## Results

### Overview

Smartphone data were available for 2010 participants aged 18 years and older. Of these, 34 (1.7%) participants were excluded from the analyses due to low or unstable sampling frequency, resulting in 1976 evaluable participants using 206 different, or unique, device models (1614 participants with iOS devices and 362 participants with Android devices). Details on the participant, the devices they used, and their demographics are provided in Figure 1 and Table 2.

**Figure 1.** Participant deposition and the devices they used.

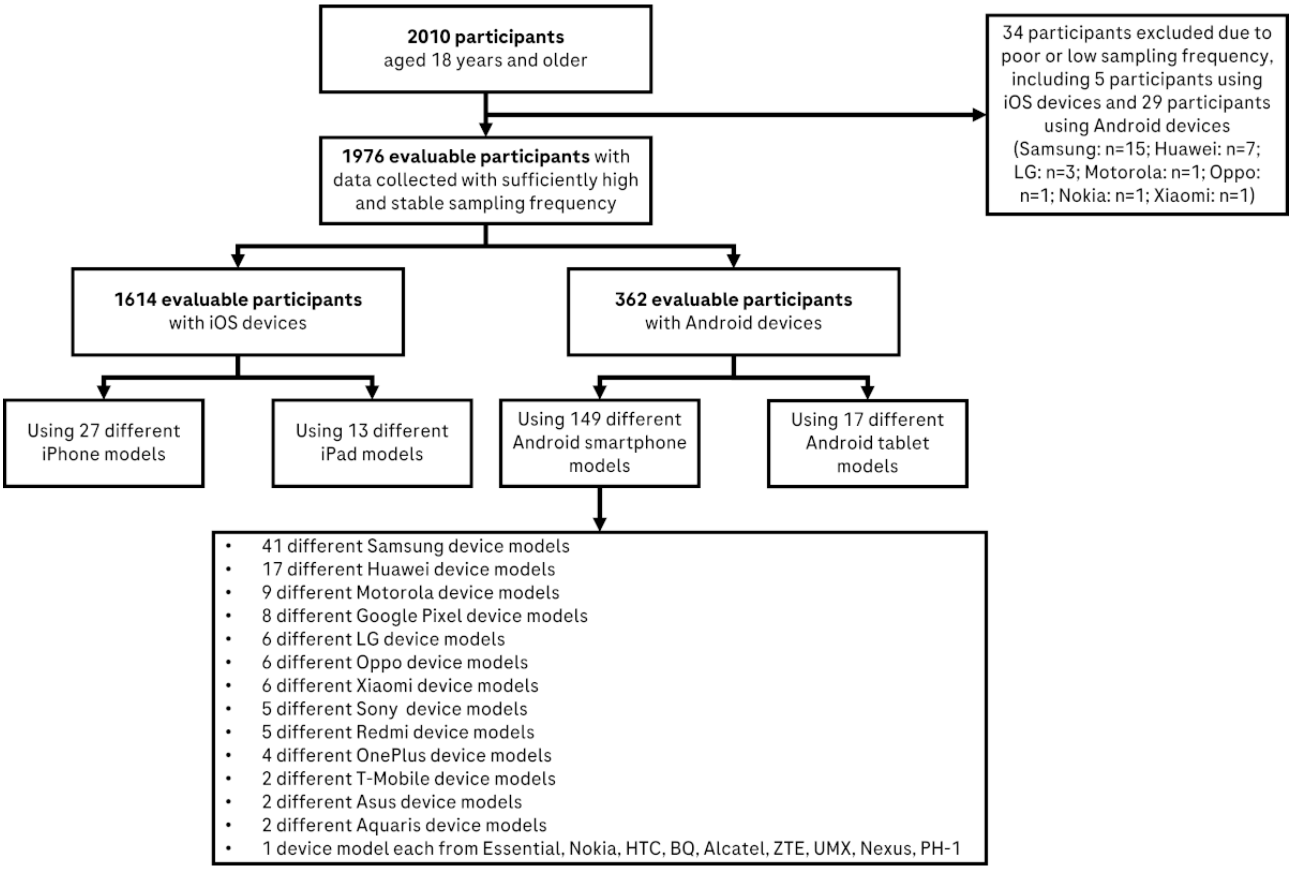

**2010 participants** aged 18 years and older
34 participants excluded due to poor or low sampling frequency, including 5 participants using iOS devices and 29 participants using Android devices (Samsung: n=15; Huawei: n=7; LG: n=3; Motorola: n=1; Oppo: n=1; Nokia: n=1; Xiaomi: n=1)
**1976 evaluable participants** with data collected with sufficiently high and stable sampling frequency
**1614 evaluable participants** with iOS devices
**362 evaluable participants** with Android devices
Using 27 different iPhone models
Using 13 different iPad models
Using 149 different Android smartphone models
Using 17 different Android tablet models
41 different Samsung device models
17 different Huawei device models
9 different Motorola device models
8 different Google Pixel device models
6 different LG device models
6 different Oppo device models
6 different Xiaomi device models
5 different Sony device models
5 different Redmi device models
4 different OnePlus device models
2 different T-Mobile device models
2 different Asus device models
2 different Aquaris device models
1 device model each from Essential, Nokia, HTC, BQ, Alcatel, ZTE, UMX, Nexus, PH-1

**Table 2.** Baseline demographics and disease characteristics.

| Variable | All (n=1976) | MS[a] (n=1140) | Non-MS (n=836) |
|---|---|---|---|
| Age (years), mean (SD) | 43.1 (12.5) | 45.5 (12.0) | 39.8 (12.4) |
| Female, n (%) | 1243 (62.9) | 838 (73.5) | 405 (48.4) |
| Participants with iOS devices, n (%) | 1614 (81.7) | 879 (77.1) | 735 (87.9) |
| **Active tests, mean (SD)** | | | |
| IPS[b] (correct responses, n) | 43.8 (11.9) | 43.2 (10.7) | 45.0 (13.7) |
| IPS DD[c] (correct responses, n) | 17.9 (3.6) | 17.7 (3.4) | 18.3 (4.0) |
| PT[d] (successful pinches, n) | 30.8 (13.5) | 28.1 (13.0) | 34.6 (13.5) |
| SBT[e] (sway path, m/$s^2$) | 21.8 (21.9) | 22.7 (22.0) | 20.5 (21.6) |
| UTT[f] (turn speed, rad/s) | 1.40 (0.35) | 1.37 (0.34) | 1.43 (0.37) |
| 2MWT[g] (steps, n) | 190.2 (38.2) | 191.2 (38.1) | 187.8 (38.2) |

[a] MS: multiple sclerosis.

[b] IPS: Information Processing Speed.

[c] IPS DD: Information Processing Speed Digit-Digit.

[d] PT: Pinching Test.

[e] SBT: Static Balance Test.

[f] UTT: U-Turn Test.

[g] 2MWT: 2-Minute Walk Test.

## Quality Checks

Devices running iOS devices were less prone to low or unstable sampling frequency than Android devices, although the overall number of participants that were excluded for failing to meet the sampling frequency requirements was small. In total, 5 participants with iOS devices and 29 participants with Android devices (15 participants with Samsung; 7 participants with Huawei; 3 participants with LG; and 1 participant each with Motorola, Oppo, Nokia, and Xiaomi devices) were excluded. Of all tests performed during the study, 95.9% (91,002/94,925 tests) were recorded with a sufficiently high and stable sampling frequency.

Across the entire study duration, the proportion of active tests executed in accordance with the test instructions, that is, which passed the criteria defined by the quality control flags, varied from approximately 61.9% to 99.1% (IPS: 6436/6985 [92.1%]; IPS DD: 6278/6609 [95.0%]; PT: 31235/31508 [99.1%]; SBT: 23696/26954 [87.9%]; UTT: 14985/24219 [61.9%]; 2MWT: 18055/21483 [84.0%]). However, no participant was excluded from the analysis for failing to meet these criteria as each participant performed at least one entire test run in accordance with the provided instructions (for each participant, the first valid test from each active test was used for the measurement equivalence analysis). Nonetheless, this highlights that large observational remote assessment studies such as Floodlight Open require checks that provide objective measures for assessing data quality in accordance with the test instructions of self-administered assessments [15].

## Measurement Equivalence

The measurement equivalence analysis was assessed after adjusting for age, sex, and self-reported disease status. Overall, results show no indication that any of the 6 active tests are associated with a systematic measurement nonequivalence. The findings were consistent across OS platforms, OS versions, and device models, although the findings for the device models were more variable given the smaller subgroups. The effect sizes of the difference between subgroups were mostly very small (effect size <0.20) or small (effect size <0.5). For each observed subgroup, the mean differences were considerably smaller than one SD of the respective reference group, and most mean differences were within the 95% CI obtained from the permutation test (Tables 3-5; Table S1 in Multimedia Appendix 1).

When comparing iOS devices with Android devices, percent differences in adjusted test scores between iOS devices and Android devices were <5% on the IPS, IPS DD, SBT, UTT, and 2MWT and ≤10% on the PT (Table 3). Permutation testing revealed a statistically significant, but small difference between iOS and Android on the PT (percent difference: 8.0%, effect size: 0.24, *P*<.001; Figure 2; Table 3).

Similar results were observed when comparing the 5 iOS versions 11, 12, 13, 14, and 15 (Figure 3A). Across all 5 iOS versions, the percent differences were mostly <5% on the IPS and IPS DD, PT, UTT, and 2MWT, with only iOS 15 showing percent differences greater than 5% on the IPS (percent difference: 5.3%) and IPS DD (percent difference: 5.2%) (Table 4). Larger differences were observed on the SBT, with percent differences ranging from 1.3% to 18.3%. Permutation testing revealed statistically significant differences only for iO12 on the IPS (percent difference: 3.0%, effect size: 0.17, $P_{unadjusted}$=.02) and IPS DD (percent difference: 2.5%, effect size: 0.19, $P_{unadjusted}$=.01). However, the observed mean differences of 1.9 (IPS) and 0.6 (IPS DD) correct responses were within the 95% CI obtained from the permutation test (IPS: [0.0,1.9]; IPS DD: [0.0,0.6]), and these differences were no longer statistically significant after correcting for multiple comparisons ($P_{adjusted}$=.11 and .07, respectively). Differences of similar magnitude were observed when comparing Android versions 8, 9, and 10 (Figure 3B). Across these OS versions, percent differences were mostly ≤5% on the IPS, IPS DD, PT (only Android 8 showed a larger percent difference of 5.1%), UTT, and 2MWT; and <30% on the SBT (Table 5). None of these differences reached statistical significance (all $P_{unadjusted}$=.06-.96, all $P_{adjusted}$=.19-.96).

Next, we evaluated whether a device model shows an out-of-distribution performance compared with the null hypothesis of all device models being from the same distribution (Table S1 and Figure S1 in Multimedia Appendix 1). Only device models used by at least 20 participants were included in this analysis, resulting in 17 different models being evaluated. Percent differences were mostly ≤5% on the IPS, IPS DD, UTT (only the iPhone 7 Plus showed a larger percent difference of 5.6%) and 2MWT; and <20% on the PT and SBT (Table S1 in Multimedia Appendix 1). Unadjusted permutation testing revealed that most device models did not show a statistically significant difference (IPS: 12/13 models, IPS DD: all models, PT: 12/16 models, SBT: 12/14 models, UTT: all models, 2MWT: all models). Of note, the statistically significant differences ($P_{unadjusted}$<.001-.04) were not associated with any particular device model. After adjusting for multiple comparisons, statistically significant differences were observed only on the PT (iPhone SE: $P_{adjusted}$=.02; iPhone X Global: $P_{adjusted}$=.03; iPhone X GSM: $P_{adjusted}$≤.001; iPhone 11: $P_{adjusted}$=.03). In all 4 instances, effect sizes ranged from 0.38 to 0.50, and percent differences from 11.4% to 15.1%. Furthermore, the observed mean differences were outside the 95% CI obtained from the permutation test (Table S1 in Multimedia Appendix 1).

The sensitivity analysis, which used the median difference rather than the mean difference as the test statistic for the permutation test, revealed similar findings across OS platforms, OS versions, and device models (Tables S2-S5 in Multimedia Appendix 1).

Screen size had a limited impact on the measurement equivalence properties. No significant differences were observed between smartphones (n=1788, of which 1556 [87.0%] were iOS devices) and tablets (n=89, of which 58 [65.2%] were iPads running iOS) on the gait and balance tests (ie, SBT, UTT and 2MWT, all $P$=.27-.75). Statistically significant differences with small effect sizes between smartphones and tablets were observed on the IPS (percent difference: 11.9%, effect size: 0.45, $P$=.001), IPS DD (percent difference: 7.3%, effect size: 0.37, $P$=.022), and PT (percent difference: 18.1%, effect size: 0.38, $P$=.002).

**Table 3.** Absolute and percent mean differences across operating system platforms (iOS vs Android).

| Active test | iOS | | Android | | Mean difference from Android devices | | | Permutation test | |
|---|---|---|---|---|---|---|---|---|---|
| | n | Mean (SD) | n | Mean (SD) | Absolute difference | Percent difference | Effect size | 95% CI | $P_{unadjusted}$ |
| IPS[a] (correct responses, n) | 734 | 63.1 (11.2) | 341 | 61.8 (10.3) | 1.3 | 2.1 | 0.12 | 0.0-1.6 | .07 |
| IPS DD[b] (correct responses, n) | 714 | 24.3 (3.4) | 333 | 24.2 (3.1) | 0.1 | 0.5 | 0.04 | 0.0-0.5 | .60 |
| PT[c] (successful pinches, n) | 1233 | 41.5 (12.7) | 338 | 38.4 (12.1) | 3.1 | 8.0 | 0.24 | 0.0-1.7 | <.001 |
| SBT[d] (sway path, m/s$^2$) | 1313 | 28.2 (20.7) | 126 | 28.4 (20.8) | 0.2 | 0.7 | 0.01 | 0.1-4.3 | .92 |
| UTT[e] (turn speed, rad/s) | 975 | 1.59 (0.34) | 219 | 1.59 (0.37) | 0.01 | 0.4 | 0.02 | 0.00-0.06 | .80 |
| 2MWT[f] (steps, n) | 619 | 195.9 (38.0) | 168 | 191.9 (38.2) | 4.0 | 2.1 | 0.11 | 0.1-7.5 | .23 |

[a] IPS: Information Processing Speed.

[b] IPS DD: Information Processing Speed Digit-Digit.

[c] PT: Pinching Test.

[d] SBT: Static Balance Test.

[e] UTT: U-Turn Test.

[f] 2MWT: 2-Minute Walk Test.

**Table 4.** Absolute and percent differences across iOS versions.

| Active test | iOS subgroup | | Reference group[a], mean (SD) | Mean difference from reference group | | | Permutation test | | |
|---|---|---|---|---|---|---|---|---|---|
| | n | Mean (SD) | | Absolute difference | Percent difference | Effect size | 95% CI | $P_{\text{unadjusted}}$ | $P_{\text{adjusted}}$[b] |
| **IPS[c] (correct responses, n)** | | | | | | | | | |
| iOS 11 | 84 | 62.9 (12.8) | 63.1 (11.0) | 0.2 | 0.4 | 0.02 | 0.0-2.9 | .86 | .87 |
| iOS 12 | 304 | 62.0 (11.5) | 63.9 (11.0) | 1.9 | 3.0 | 0.17 | 0.0-1.9 | .02 | .11 |
| iOS 13 | 199 | 63.7 (11.8) | 62.9 (11.0) | 0.8 | 1.2 | 0.07 | 0.0-2.1 | .40 | .50 |
| iOS 14 | 118 | 64.5 (8.5) | 62.8 (11.7) | 1.7 | 2.7 | 0.15 | 0.0-2.5 | .12 | .27 |
| iOS 15 | 23 | 66.3 (9.0) | 63.0 (11.3) | 3.3 | 5.3 | 0.30 | 0.1-5.3 | .16 | .27 |
| **IPS DD[d] (correct responses, n)** | | | | | | | | | |
| iOS 11 | 83 | 24.2 (3.9) | 24.3 (3.3) | 0.1 | 0.5 | 0.03 | 0.0-0.9 | .78 | .78 |
| iOS 12 | 296 | 23.9 (3.5) | 24.5 (3.3) | 0.6 | 2.5 | 0.19 | 0.0-0.6 | .01 | .07 |
| iOS 13 | 190 | 24.5 (3.3) | 24.2 (3.4) | 0.2 | 0.9 | 0.07 | 0.0-0.6 | .43 | .53 |
| iOS 14 | 117 | 24.8 (2.7) | 24.2 (3.5) | 0.6 | 2.4 | 0.17 | 0.0-0.8 | .09 | .16 |
| iOS 15 | 23 | 25.5 (3.1) | 24.2 (3.4) | 1.3 | 5.2 | 0.38 | 0.0-1.6 | .08 | .16 |
| **PT[e] (successful pinches, n)** | | | | | | | | | |
| iOS 11 | 38 | 41.2 (11.7) | 41.5 (12.7) | 0.3 | 0.7 | 0.02 | 0.1-4.7 | .89 | .99 |
| iOS 12 | 568 | 41.6 (12.0) | 41.3 (13.3) | 0.3 | 0.7 | 0.02 | 0.0-1.6 | .71 | .99 |
| iOS 13 | 399 | 41.5 (13.6) | 41.5 (12.3) | 0.0 | 0.0 | 0.00 | 0.0-1.8 | .99 | .99 |
| iOS 14 | 185 | 40.9 (12.8) | 41.6 (12.7) | 0.7 | 1.7 | 0.06 | 0.0-2.3 | .48 | .99 |
| iOS 15 | 38 | 42.6 (15.0) | 41.4 (12.6) | 1.2 | 2.9 | 0.09 | 0.1-4.8 | .58 | .99 |
| **SBT[f] (sway path, m/s$^2$)** | | | | | | | | | |
| iOS 11 | 181 | 29.1 (21.5) | 28.1 (20.5) | 1.0 | 3.7 | 0.05 | 0.0-3.7 | .54 | .67 |
| iOS 12 | 578 | 29.1 (21.0) | 27.5 (20.4) | 1.6 | 5.7 | 0.08 | 0.0-2.6 | .17 | .29 |
| iOS 13 | 348 | 26.7 (20.0) | 28.7 (20.9) | 2.0 | 7.1 | 0.10 | 0.0-2.9 | .11 | .29 |
| iOS 14 | 163 | 28.5 (21.7) | 28.2 (20.5) | 0.4 | 1.3 | 0.02 | 0.1-3.9 | .83 | .83 |
| iOS 15 | 38 | 23.2 (12.0) | 28.3 (20.9) | 5.2 | 18.3 | 0.25 | 0.1-7.6 | .13 | .29 |
| **UTT[g] (turn speed, rad/s)** | | | | | | | | | |
| iOS 11 | 121 | 1.58 (0.35) | 1.59 (0.34) | 0.01 | 0.4 | 0.02 | 0.00-0.08 | .85 | 1.0 |
| iOS 12 | 409 | 1.57 (0.33) | 1.60 (0.35) | 0.03 | 2.0 | 0.09 | 0.00-0.05 | .15 | .38 |
| iOS 13 | 279 | 1.62 (0.38) | 1.58 (0.33) | 0.04 | 2.7 | 0.12 | 0.00-0.06 | .08 | .38 |
| iOS 14 | 134 | 1.59 (0.31) | 1.59 (0.35) | 0.00 | 0.0 | 0.00 | 0.00-0.07 | 1.0 | 1.0 |
| iOS 15 | 29 | 1.59 (0.31) | 1.59 (0.35) | 0.00 | 0.2 | 0.01 | 0.00-0.15 | .96 | 1.0 |
| **2MWT[h] (steps, n)** | | | | | | | | | |
| iOS 11 | 67 | 204.1 (33.7) | 194.8 (38.4) | 9.3 | 4.8 | 0.25 | 0.2-10.9 | .057 | .16 |
| iOS 12 | 241 | 192.2 (42.1) | 198.1 (35.0) | 5.9 | 3.0 | 0.15 | 0.1-7.1 | .06 | .16 |
| iOS 13 | 179 | 197.7 (34.1) | 195.0 (39.5) | 2.6 | 1.4 | 0.07 | 0.1-7.6 | .43 | .45 |
| iOS 14 | 96 | 198.4 (36.1) | 195.3 (38.4) | 3.1 | 1.6 | 0.08 | 0.1-9.5 | .45 | .45 |
| iOS 15 | 32 | 186.9 (38.7) | 196.3 (38.0 | 9.4 | 4.8 | 0.25 | 0.2-15.7 | .17 | .29 |

[a] The reference group consists of all other iOS devices pooled together.

[b] *P* values were adjusted for multiple comparisons with false discovery rate (FDR) correction using the Benjamini-Hochberg method.

[c]IPS: Information Processing Speed.

[d]IPS DD: Information Processing Speed Digit-Digit.

[e]PT: Pinching Test.

[f]SBT: Static Balance Test.

[g]UTT: U-Turn Test.

[h]2MWT: 2-Minute Walk Test.

**Table 5.** Absolute and percent differences across Android versions.

| Active test | Android subgroup | | Reference group[a], mean (SD) | Mean difference from reference group | | | Permutation Test | | |
|---|---|---|---|---|---|---|---|---|---|
| | n | Mean (SD) | | Absolute difference | Percent difference | Effect size | 95% CI | *P* unadjusted | *P* adjusted [b] |
| **IPS[c] (correct responses, n)** | | | | | | | | | |
| Android 8 | 62 | 62.4 (11.5) | 61.8 (10.5) | 0.6 | 0.9 | 0.05 | 0.1-3.6 | .71 | .90 |
| Android 9 | 114 | 61.8 (10.5) | 62.1 (10.9) | 0.3 | 0.5 | 0.03 | 0.0-3.0 | .83 | .90 |
| Android 10 | 75 | 61.9 (10.6) | 62.0 (10.8) | 0.2 | 0.3 | 0.02 | 0.0-3.3 | .90 | .90 |
| **IPS DD[d] (correct responses, n)** | | | | | | | | | |
| Android 8 | 62 | 24.4 (2.5) | 24.2 (3.5) | 0.3 | 1.1 | 0.09 | 0.0-1.1 | .57 | .85 |
| Android 9 | 109 | 24.3 (3.6) | 24.2 (3.0) | 0.1 | 0.3 | 0.03 | 0.0-0.9 | .85 | .85 |
| Android 10 | 74 | 24.0 (3.4) | 24.3 (3.2) | 0.3 | 1.4 | 0.11 | 0.0-1.0 | .46 | .85 |
| **PT[e] (successful pinches, n)** | | | | | | | | | |
| Android 8 | 61 | 40.0 (12.0) | 38.1 (12.0) | 1.9 | 5.1 | 0.16 | 0.1-3.9 | .27 | .57 |
| Android 9 | 113 | 37.8 (11.3) | 39.2 (12.6) | 1.3 | 3.4 | 0.11 | 0.1-3.5 | .38 | .57 |
| Android 10 | 74 | 38.5 (13.3) | 38.6 (11.5) | 0.1 | 0.3 | 0.01 | 0.1-3.7 | .94 | .94 |
| **SBT[f] (sway path, m/s$^2$)** | | | | | | | | | |
| Android 8 | 23 | 19.8 (7.8) | 28.2 (20.6) | 8.4 | 29.8 | 0.46 | 0.2-10.0 | .06 | .19 |
| Android 9 | 38 | 28.7 (21.7) | 23.8 (15.3) | 4.9 | 20.6 | 0.27 | 0.1-9.1 | .23 | .35 |
| Android 10 | 24 | 27.5 (19.8) | 25.3 (18.0) | 2.2 | 8.7 | 0.12 | 0.1-9.9 | .63 | .64 |
| **UTT[g] (turn speed, rad/s)** | | | | | | | | | |
| Android 8 | 38 | 1.59 (0.37) | 1.59 (0.32) | 0.00 | 0.2 | 0.01 | 0.00-0.14 | .96 | .96 |
| Android 9 | 72 | 1.62 (0.34) | 1.57 (0.32) | 0.05 | 3.1 | 0.15 | 0.00-0.12 | .36 | .54 |
| Android 10 | 55 | 1.55 (0.30) | 1.60 (0.35) | 0.05 | 3.2 | 0.15 | 0.00-0.12 | .35 | .54 |
| **2MWT[h] (steps, n)** | | | | | | | | | |
| Android 8 | 33 | 191.1 (33.8) | 191.8 (42.6) | 0.7 | 0.4 | 0.02 | 0.3-18.3 | .94 | .94 |
| Android 9 | 57 | 195.7 (41.0) | 188.3 (40.0) | 7.3 | 3.9 | 0.18 | 0.2-16.1 | .31 | .46 |
| Android 10 | 39 | 186.0 (45.3) | 194.0 (38.1) | 7.9 | 4.1 | 0.20 | 0.3-17.3 | .30 | .46 |

[a]The reference group consists of all other Android devices pooled together.

[b]*P* values were adjusted for multiple comparisons with false discovery rate (FDR) correction using the Benjamini-Hochberg method.

[c]IPS: Information Processing Speed.

[d]IPS DD: Information Processing Speed Digit-Digit.

[e]PT: Pinching Test.

[f]SBT: Static Balance Test.

[g]UTT: U-Turn Test.

[h]2MWT: 2-Minute Walk Test.

**Figure 2.** Measurement equivalence by operating system (OS) platform. Permutation testing showed no evidence of a systematic lack of equivalence between the 2 OS platforms—iOS and Android. A small but statistically significant difference was only observed on the PT. Differences between groups as well as effect sizes are provided in Table 3. Brackets indicate the sample size. ****P*<.001. IPS: Information Processing Speed; IPS DD: Information Processing Speed Digit-Digit; PT: Pinching Test; SBT: Static Balance Test; UTT: U-Turn Test; 2MWT: 2-Minute Walk Test.

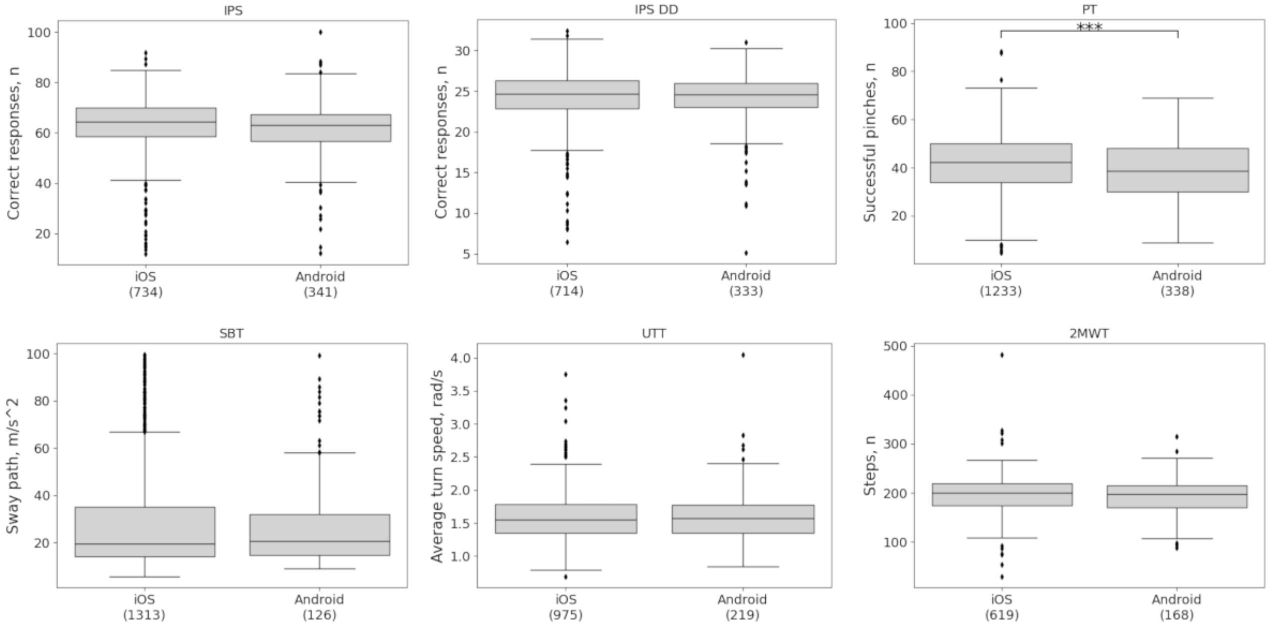
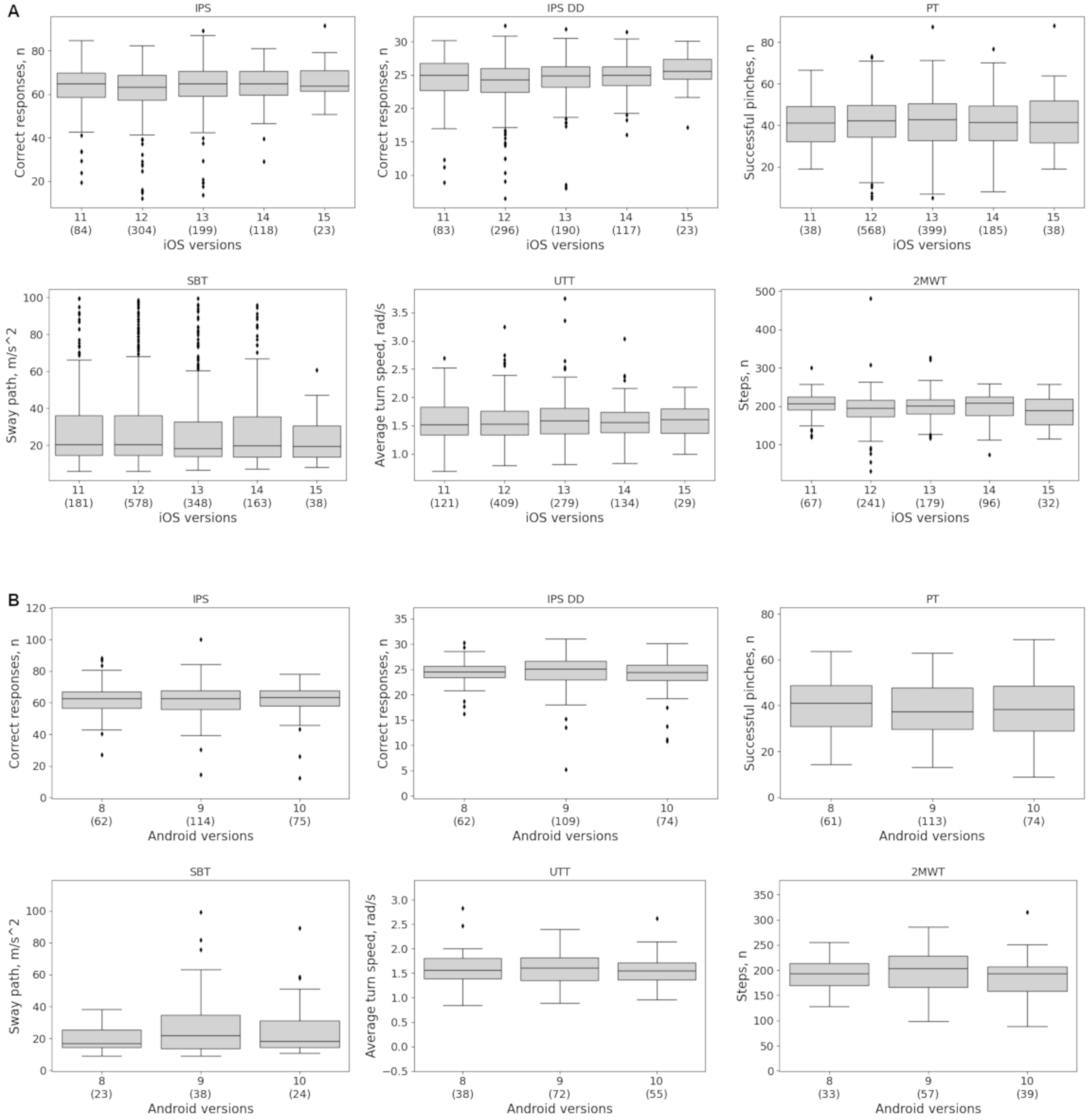


**Figure 3.** Measurement equivalence by (A) iOS version and (B) Android version. Permutation testing revealed no evidence of a systematic lack of equivalence. No statistically significant differences were observed across operating system versions after correcting for multiple comparisons (all $P_{adjusted}$>.05). Absolute and percent differences as well as effect sizes are provided in Tables 4 and 5. Brackets indicate the sample size. IPS: Information Processing Speed; IPS DD: Information Processing Speed Digit-Digit; PT: Pinching Test; SBT: Static Balance Test; UTT: U-Turn Test; 2MWT: 2-Minute Walk Test.



## Discussion

### Principal Findings

Ensuring measurement equivalence across different device models is necessary to conduct a new form of large-scale, global, and cost-effective clinical research on digital outcomes in a BYOD setting. In this analysis, no evidence was found for a systematic lack of measurement equivalence across devices for the 6 smartphone sensor–based active tests considered. The key learnings are summarized in the Textbox 1.

Measurement nonequivalence can in theory arise from various sources. For example, differences in inertial measurement unit (IMU) sensors can potentially contribute to differences in measurements obtained between devices [35]. However, our results suggest that differences in IMU sensors did not impact the features derived from our active tests as no statistically significant differences were observed on tests that rely on IMU sensors (ie, the SBT, UTT, and 2MWT).

**Textbox 1.** Key learnings.

- The broad, multinational cohort of 1976 participants and 206 different device models allowed us to study measurement equivalence in a naturalistic environment, reflective of a real-world setting.
- No evidence for the systematic lack of measurement equivalence was found; the features derived from our smartphone-based assessments were by large robust against differences in inertial measurement unit (IMU) sensors, sampling frequency, and device latency.
- Designing features to be robust against cumulative device latency effects can help ensure measurement equivalence across operating system (OS) platforms, OS versions, and device models.
- Objective criteria that assess whether a self-administered digital test was taken as instructed and recorded with a sufficiently high and stable sampling frequency can help with evaluating the quality of the data collected in an unsupervised setting and, hence, informing whether the data should be included in further analyses.

Another factor that could contribute to measurement nonequivalence is differences in sampling frequency or sampling frequency heterogeneity [35]. We, therefore, excluded any tests that were recorded with either too low a sampling frequency or with an unstable sampling frequency as the sampling frequency can vary from test execution to execution, even with the same device model. Nonetheless, the sampling frequency requirements were met in most cases even if the computational load was temporarily increased due to the simultaneous activity of other apps installed on the participants’ own smartphone devices as can be expected in a BYOD setting. Of the 2010 participants considered for the analyses, only 34 participants were excluded from the analysis for the lack of a sufficiently high and stable sample frequency.

A third factor that could impact measurement equivalence is device latency. For touchscreen-based tests, it was estimated that latencies associated with displaying a new visual cue on the screen and registering a touchscreen input by the user could account for a variance in response time of up to 100 ms, with Android devices tending to show higher latencies than iOS devices [41]. We estimated that such latencies could, in the worst case, account for a difference of up to 3 pinches on the PT (100 ms latency per pinched tomato × an average 30.8 pinched tomatoes during a PT [Table 2] = 3.08 seconds latency during a PT; with approximately 1 pinched tomato per second, this cumulative latency may thus explain a difference of up to 3 successful pinches). This is comparable to the 3.1 fewer successful pinches observed on Android devices than on iOS devices. While other factors cannot be ruled out, it is plausible that the difference of 3.1 successful pinches between Android and iOS devices observed in this study could be explained by differences in device latencies. Developing features that are not affected by such cumulative latency effects could improve the measurement equivalence between iOS and Android devices. For the PT, a possible feature is a double touch asynchrony, which was developed for other studies that deployed the Floodlight technology as a measure of finger coordination. It measures the duration of the gap between the first and second fingers touching the touch screen. In a previous analysis, it was reliable, correlated with clinician-administered measures of upper extremity impairment and MS-related disability, and differentiated between people with different levels of disability [42]. Unlike the number of successful pinches, this feature is by design not affected by cumulative latency effects as it measures the average time between the first and second finger touching the touch screen during a single PT.

Our study also highlighted the need for objective measures of data quality and accordance with the test instructions. Across the 6 active tests considered, approximately 61.9% through 99.1% of all test attempts were considered valid. A higher proportion of tests passing the quality-check flags have been previously reported when using less stringent criteria [33]. Ensuring that the tests are executed in accordance with their instructions remains a challenge with fully remote digital health studies which do not involve direct oversight of test performance and feedback mechanisms with the participants. In-clinic visits, for example, could be used to onboard study participants in person and explain the tests in more detail, consequently increasing the proportion of active tests executed in accordance with the test instructions. In fact, we previously reported for a separate study, which included regular clinic visits, that up to 99% of active test attempts were considered valid [15]. Furthermore, digital health studies that include clinic visits, such as [15], are more likely than fully remote studies to enroll participants who are more motivated. This becomes more noticeable the longer the test duration is and the smaller the subgroups that are compared with each other. As a result, we implemented in this study additional quality control flags to ensure sufficient evaluable data were collected during the 2 active tests with the longest test duration, the UTT and 2MWT (Table 1).

## Limitations

There are some limitations associated with our study. First, the observational nature of our study in a naturalistic environment implies that confounding factors may have affected the comparisons despite the adjustments we performed. For instance, we could not adjust for disease severity, as this information is not available in this fully remote, digital-only study. This might partially explain the larger percent differences observed in smaller subgroups. However, given the small overall differences among the investigated subgroups, the residual confounding effects are plausibly small. While statistical tests did not indicate that the measurements obtained across devices are systematically nonequivalent, we cannot rule out that this finding may be due to high variability, different average disease severity, or, for a few comparisons, a small sample size. A lack of sensitivity of the Floodlight tests, however, can be excluded as a reason for the similarity of the results as these tests have shown the ability to differentiate people with MS from healthy controls or between people with MS with different levels of impairment [15,42,43].

Second, the earlier release of the iOS version of the Floodlight Open app has resulted in a larger number of iOS participants compared with Android participants. Furthermore, the larger choice of different Android device models than iOS device models meant that there were fewer participants per Android device model, resulting in an unbalanced data set for the comparison of measurement equivalence by device model.

Finally, the analyses presented here are retrospective as the study was not purposely designed to assess equivalence or reliability. Ideally, device equivalence studies with purpose-built study designs have each participant perform the active tests multiple times using the same set of devices and assess equivalence across devices using appropriate statistical tests with prespecified equivalence margins. As an alternative, equivalence can be assessed with bench testing using robots able to cover the range of feature values typically observed in the population of interest. Nonetheless, the preliminary findings of our analyses can inform the design of future device equivalence studies and the equivalence margins to assess equivalence.

### Comparison With Prior Work

Recent studies compared smartphone-based measurements of gait, balance, or physical activity [44-48], cognitive functioning [49,50], or sleep [51]. However, these studies typically compared the smartphone-based measurements and measurements obtained with other devices with gold-standard measures such as manual step count [46,47] or smartphone-based measurements with measurements obtained with other device types [44,45,51]. This makes it challenging to separate the potential impact arising from differences in the device or sensor specifications from those arising from differences in the algorithms used to extract the features. A small number of studies assessed device equivalence using the same feature extraction algorithm [35,49]. For example, van Oirschot et al [49] investigated their smartphone-based Symbol Digit Modalities Test on devices running either iOS or Android and found no differences between the 2 OS platforms, which is in line with our findings on the IPS. Additionally, Ena et al [52] identified a number of features extracted from their smartphone-based mobility tests that showed in a prospective study reliability across iOS and Android devices.

Strengths of this remote, digital-only, unsupervised, BYOD study include the comprehensive approach to assess measurement equivalence across different smartphone devices. Unlike other studies, we investigated measurement equivalence on 3 levels by separately looking at the 2 OS platforms iOS and Android, different iOS and Android versions, and device models. The large sample size of 1976 study participants and the wide range of smartphone device models (n=206) used add to the strength of our study. Finally, we used objective measures to ensure data quality and accordance with test instructions.

### Conclusions

In this study, we investigated a key criterion for the deployment of 6 smartphone sensor–based active tests assessing cognition, hand motor function, and gait and balance in a BYOD setting. Using data collected in Floodlight Open, we found no evidence for a substantial and consistent lack of measurement equivalence across different OS platforms, OS versions, and smartphone device models for features derived from these active tests. These features were largely unaffected by differences in IMU and touchscreen sensors and were also not impacted by the activity of other apps installed on the participants' own smartphone devices, as evidenced by the sufficiently high and stable sampling frequency in a vast majority of participants. These findings are compatible with the use of smartphone-based tests in a BYOD setting and will inform future studies with purpose-built study designs that will further investigate the device equivalence properties of smartphone-based assessments.

## Acknowledgments

We would like to thank all the Roche Floodlight Open team members, particularly Rita Freitas, Laura Alberti, Philipp D'Alfonso, Iris Shih, Grégoire Pointeau, and all the participants and the researchers who expressed interest in this study. This research was funded by F. Hoffmann-La Roche Ltd, Basel, Switzerland. Editorial and writing support was provided by Sven Holm, PhD, contractor for F. Hoffmann-La Roche Ltd. The authors had full editorial control of the manuscript and provided their final approval of all content.

## Data Availability

Up-to-date details on Roche's Global Policy on Sharing of Clinical Study Information and how to request access to related clinical study documents are available in [53]. Request for the data underlying this publication requires a detailed, hypothesis-driven, statistical analysis plan that is collaboratively developed by the requestor and company subject matter experts. Such requests should be directed to dbm.datarequest@roche.com for consideration. Anonymized records for individual patients across more than one data source external to Roche cannot, and should not, be linked due to a potential increase in risk of patient reidentification.

## Conflicts of Interest

LK, LC, FL, and AG are employees of F. Hoffmann-La Roche Ltd. FD was an employee of F. Hoffmann-La Roche Ltd. during the completion of the work related to this manuscript. FD is now an employee of Novartis (Basel, Switzerland), which was not in any way associated with this study. CB was a contractor for F. Hoffmann-La Roche Ltd. during the completion of the work related to this manuscript. CB is now with Limites Medical Research Ltd., which was not in any way associated with this study. ML is a consultant for F. Hoffmann-La Roche Ltd via Inovigate.

## Multimedia Appendix 1

Absolute and percent mean difference by device model and sensitivity analysis results.
[DOCX File , 387 KB-Multimedia Appendix 1]

## References

1. Reich DS, Lucchinetti CF, Calabresi PA. Multiple sclerosis. N Engl J Med. 2018;378(2):169-180. [FREE Full text] [doi: 10.1056/NEJMra1401483] [Medline: 29320652]
2. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: insights from the Atlas of MS, third edition. Mult Scler. 2020;26(14):1816-1821. [FREE Full text] [doi: 10.1177/1352458520970841] [Medline: 33174475]
3. Hobart J, Bowen A, Pepper G, Crofts H, Eberhard L, Berger T, et al. International consensus on quality standards for brain health-focused care in multiple sclerosis. Mult Scler. 2019;25(13):1809-1818. [FREE Full text] [doi: 10.1177/1352458518809326] [Medline: 30381987]
4. Rae-Grant A, Bennett A, Sanders AE, Phipps M, Cheng E, Bever C. Quality improvement in neurology: multiple sclerosis quality measures: executive summary. Neurology. 2015;85(21):1904-1908. [FREE Full text] [doi: 10.1212/WNL.0000000000001965] [Medline: 26333795]
5. Clay I. Impact of digital technologies on novel endpoint capture in clinical trials. Clin Pharmacol Ther. 2017;102(6):912-913. [doi: 10.1002/cpt.866] [Medline: 29027665]
6. Boukhvalova AK, Fan O, Weideman AM, Harris T, Kowalczyk E, Pham L, et al. Smartphone level test measures disability in sveral neurological domains for patients with multiple sclerosis. Front Neurol. 2019;10:358. [FREE Full text] [doi: 10.3389/fneur.2019.00358] [Medline: 31191424]
7. Maillart E, Labauge P, Cohen M, Maarouf A, Vukusic S, Donzé C, et al. MSCopilot, a new multiple sclerosis self-assessment digital solution: results of a comparative study versus standard tests. Eur J Neurol. 2020;27(3):429-436. [doi: 10.1111/ene.14091] [Medline: 31538396]
8. Block VJ, Bove R, Zhao C, Garcha P, Graves J, Romeo AR, et al. Association of continuous assessment of step count by remote monitoring with disability progression among adults with multiple sclerosis. JAMA Netw Open. 2019;2(3):e190570. [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0570] [Medline: 30874777]
9. Pham L, Harris T, Varosanec M, Morgan V, Kosa P, Bielekova B. Smartphone-based symbol-digit modalities test reliably captures brain damage in multiple sclerosis. NPJ Digit Med. 2021;4(1):36. [FREE Full text] [doi: 10.1038/s41746-021-00401-y] [Medline: 33627777]
10. Midaglia L, Mulero P, Montalban X, Graves J, Hauser SL, Julian L, et al. Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: nonrandomized interventional feasibility study. J Med Internet Res. 2019;21(8):e14863. [FREE Full text] [doi: 10.2196/14863] [Medline: 31471961]
11. van der Walt A, Butzkueven H, Shin RK, Midaglia L, Capezzuto L, Lindemann M, et al. Developing a digital solution for remote assessment in multiple sclerosis: from concept to software as a medical device. Brain Sci. 2021;11(9):1247. [FREE Full text] [doi: 10.3390/brainsci11091247] [Medline: 34573267]
12. Chitnis T, Glanz BI, Gonzalez C, Healy BC, Saraceno TJ, Sattarnezhad N, et al. Quantifying neurologic disease using biosensor measurements in-clinic and in free-living settings in multiple sclerosis. NPJ Digit Med. 2019;2:123. [FREE Full text] [doi: 10.1038/s41746-019-0197-7] [Medline: 31840094]
13. Angelini L, Hodgkinson W, Smith C, Dodd JM, Sharrack B, Mazzà C, et al. Wearable sensors can reliably quantify gait alterations associated with disability in people with progressive multiple sclerosis in a clinical setting. J Neurol. 2020;267(10):2897-2909. [FREE Full text] [doi: 10.1007/s00415-020-09928-8] [Medline: 32468119]
14. Lipsmeier F, Taylor KI, Postuma RB, Volkova-Volkmar E, Kilchenmann T, Mollenhauer B, et al. Reliability and validity of the roche PD mobile application for remote monitoring of early Parkinson's disease. Sci Rep. 2022;12(1):12081. [FREE Full text] [doi: 10.1038/s41598-022-15874-4] [Medline: 35840753]
15. Montalban X, Graves J, Midaglia L, Mulero P, Julian L, Baker M, et al. A smartphone sensor-based digital outcome assessment of multiple sclerosis. Mult Scler. 2022;28(4):654-664. [FREE Full text] [doi: 10.1177/13524585211028561] [Medline: 34259588]
16. Lam KH, Bucur IG, van Oirschot P, de Graaf F, Weda H, Strijbis E, et al. Towards individualized monitoring of cognition in multiple sclerosis in the digital era: a one-year cohort study. Mult Scler Relat Disord. 2022;60:103692. [FREE Full text] [doi: 10.1016/j.msard.2022.103692] [Medline: 35219240]
17. Creagh AP, Lipsmeier F, Lindemann M, Vos MD. Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones. Sci Rep. 2021;11(1):14301. [FREE Full text] [doi: 10.1038/s41598-021-92776-x] [Medline: 34253769]
18. Creagh AP, Simillion C, Scotland A, Lipsmeier F, Bernasconi C, Belachew S, et al. Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the draw a shape test. Physiol Meas. 2020;41(5):054002. [FREE Full text] [doi: 10.1088/1361-6579/ab8771] [Medline: 32259798]

19. Feys P, Lamers I, Francis G, Benedict R, Phillips G, LaRocca N, et al. The nine-hole peg test as a manual dexterity performance measure for multiple sclerosis. Mult Scler. 2017;23(5):711-720. [FREE Full text] [doi: 10.1177/1352458517690824] [Medline: 28206826]
20. Smith A. Symbol Digit Modalities Test: Manual. Los Angeles, CA. Western Psychological Services; 1991.
21. Motl RW, Cohen JA, Benedict R, Phillips G, LaRocca N, Hudson LD, et al. Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis. Mult Scler. 2017;23(5):704-710. [FREE Full text] [doi: 10.1177/1352458517690823] [Medline: 28206828]
22. Marziniak M, Brichetto G, Feys P, Meyding-Lamadé U, Vernon K, Meuth SG. The use of digital and remote communication technologies as a tool for multiple sclerosis management: narrative review. JMIR Rehabil Assist Technol. 2018;5(1):e5. [FREE Full text] [doi: 10.2196/rehab.7805] [Medline: 29691208]
23. Giovannoni G, Butzkueven H, Dhib-Jalbut S, Hobart J, Kobelt G, Pepper G, et al. Brain health: time matters in multiple sclerosis. Mult Scler Relat Disord. 2016;9 Suppl 1:S5-S48. [FREE Full text] [doi: 10.1016/j.msard.2016.07.003] [Medline: 27640924]
24. Affinito L, Fontanella A, Montano N, Brucato A. How physicians can empower patients with digital tools. J Public Health (Berl.). 2020;30(4):897-909. [FREE Full text] [doi: 10.1007/s10389-020-01370-4]
25. Seshadri DR, Bittel B, Browsky D, Houghtaling P, Drummond CK, Desai M, et al. Accuracy of the Apple Watch 4 to measure heart rate in patients with atrial fibrillation. IEEE J Transl Eng Health Med. 2020;8:2700204. [FREE Full text] [doi: 10.1109/JTEHM.2019.2950397] [Medline: 32128290]
26. de Redon E, Centi A. Realities of conducting digital health research: challenges to consider. Digit Health. 2019;5:2055207619869466. [FREE Full text] [doi: 10.1177/2055207619869466] [Medline: 31448129]
27. Statcounter GlobalStats. Mobile operating system market share. URL: https://gs.statcounter.com/os-market-share/mobile/worldwide [accessed 2022-01-27]
28. Byrom B, Doll H, Muehlhausen W, Flood E, Cassedy C, McDowell B, et al. Measurement equivalence of patient-reported outcome measure response scale types collected using bring your own device compared to paper and a provisioned device: results of a randomized equivalence trial. Value Health. 2018;21(5):581-589. [FREE Full text] [doi: 10.1016/j.jval.2017.10.008] [Medline: 29753356]
29. Pugliese L, Woodriff M, Crowley O, Lam V, Sohn J, Bradley S. Feasibility of the "bring your own device" model in clinical research: results from a randomized controlled pilot study of a mobile patient engagement tool. Cureus. 2016;8(3):e535. [FREE Full text] [doi: 10.7759/cureus.535] [Medline: 27096135]
30. Coons SJ, Eremenco S, Lundy JJ, O'Donohoe P, O'Gorman H, Malizia W. Capturing patient-reported outcome (PRO) data electronically: the past, present, and promise of ePRO measurement in clinical trials. Patient. 2015;8(4):301-309. [FREE Full text] [doi: 10.1007/s40271-014-0090-z] [Medline: 25300613]
31. Gwaltney C, Coons SJ, O'Donohoe P, O'Gorman H, Denomey M, Howry C, et al. "Bring Your Own Device" (BYOD): the future of field-based patient-reported outcome data collection in clinical trials? Ther Innov Regul Sci. 2015;49(6):783-791. [doi: 10.1177/2168479015609104] [Medline: 30222388]
32. Ferrar J, Griffith GJ, Skirrow C, Cashdollar N, Taptiklis N, Dobson J, et al. Developing digital tools for remote clinical research: how to evaluate the validity and practicality of active assessments in field settings. J Med Internet Res. 2021;23(6):e26004. [FREE Full text] [doi: 10.2196/26004] [Medline: 34142972]
33. Oh J, Capezzuto L, Kriara L, Schjodt-Eriksen J, van Beek J, Bernasconi C, et al. Use of smartphone-based remote assessments of multiple sclerosis in Floodlight Open, a global, prospective, open-access study. Sci Rep. 2024;14(1):122. [FREE Full text] [doi: 10.1038/s41598-023-49299-4] [Medline: 38168498]
34. Schwab P, Karlen W. A deep learning approach to diagnosing multiple sclerosis from smartphone data. IEEE J Biomed Health Inform. 2021;25(4):1284-1291. [FREE Full text] [doi: 10.1109/JBHI.2020.3021143] [Medline: 32877343]
35. Stisen A, Blunck H, Bhattacharya S, Prentow T, Kjærgaard M, Dey A. Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. 2015. Presented at: SenSys '15: The 13th ACM Conference on Embedded Network Sensor Systems; November 1-4, 2015:127-140; Seoul, South Korea. [doi: 10.1145/2809695.2809718]
36. Yang M, Zheng H, Wang H, McClean S, Harris N. Assessing the utility of smart mobile phones in gait pattern analysis. Health Technol. 2012;2(1):81-88. [doi: 10.1007/s12553-012-0021-8]
37. Liu XS. The permutation test: a simple way to test hypotheses. Nurse Res. 2024;32(2):8-13. [doi: 10.7748/nr.2024.e1920] [Medline: 38289026]
38. Assaf E, Bond RM, Cranmer SJ, Kaizar EE, Ratliff Santoro L, Shikano S, et al. Understanding the relationship between official and social information about infectious disease: experimental analysis. J Med Internet Res. 2021;23(11):e25287. [FREE Full text] [doi: 10.2196/25287] [Medline: 34817389]
39. Cohen J. Statistical Power Analysis for the Behavioral Sciences. London, United Kingdom. Routledge; 1988.
40. Sawilowsky SS. New effect size rules of thumb. J Mod App Stat Meth. 2009;8(2):597-599. [doi: 10.22237/jmasm/1257035100]
41. Nicosia J, Wang B, Aschenbrenner AJ, Sliwinski MJ, Yabiku ST, Roque NA, et al. To BYOD or not: are device latencies important for bring-your-own-device (BYOD) smartphone cognitive testing? Behav Res Methods. 2023;55(6):2800-2812. [FREE Full text] [doi: 10.3758/s13428-022-01925-1] [Medline: 35953659]

42. Graves JS, Elantkowski M, Zhang YP, Dondelinger F, Lipsmeier F, Bernasconi C, et al. Assessment of upper extremity function in multiple sclerosis: feasibility of a digital pinching test. JMIR Form Res. 2023;7:e46521. [FREE Full text] [doi: 10.2196/46521] [Medline: 37782540]
43. Graves JS, Ganzetti M, Dondelinger F, Lipsmeier F, Belachew S, Bernasconi C, et al. Preliminary validity of the draw a shape test for upper extremity assessment in multiple sclerosis. Ann Clin Transl Neurol. 2023;10(2):166-180. [FREE Full text] [doi: 10.1002/acn3.51705] [Medline: 36563127]
44. Hsieh KL, Sosnoff JJ. Smartphone accelerometry to assess postural control in individuals with multiple sclerosis. Gait Posture. 2021;84:114-119. [doi: 10.1016/j.gaitpost.2020.11.011] [Medline: 33307327]
45. Piccinini F, Martinelli G, Carbonaro A. Accuracy of mobile applications versus wearable devices in long-term step measurements. Sensors (Basel). 2020;20(21):6293. [FREE Full text] [doi: 10.3390/s20216293] [Medline: 33167361]
46. Hartung V, Sarshar M, Karle V, Shammas L, Rashid A, Roullier P, et al. Validity of consumer activity monitors and an algorithm using smartphone data for measuring steps during different activity types. Int J Environ Res Public Health. 2020;17(24):9314. [FREE Full text] [doi: 10.3390/ijerph17249314] [Medline: 33322833]
47. Balto JM, Kinnett-Hopkins DL, Motl RW. Accuracy and precision of smartphone applications and commercially available motion sensors in multiple sclerosis. Mult Scler J Exp Transl Clin. 2016;2:2055217316634754. [FREE Full text] [doi: 10.1177/2055217316634754] [Medline: 28607720]
48. Orange ST, Metcalfe JW, Liefeith A, Jordan AR. Validity of various portable devices to measure sit-to-stand velocity and power in older adults. Gait Posture. 2020;76:409-414. [doi: 10.1016/j.gaitpost.2019.12.003] [Medline: 31945676]
49. van Oirschot P, Heerings M, Wendrich K, den Teuling B, Martens MB, Jongen PJ. Symbol digit modalities test variant in a smartphone app for persons with multiple sclerosis: validation study. JMIR Mhealth Uhealth. 2020;8(10):e18160. [FREE Full text] [doi: 10.2196/18160] [Medline: 33016886]
50. Passell E, Strong RW, Rutter LA, Kim H, Scheuer L, Martini P, et al. Cognitive test scores vary with choice of personal digital device. Behav Res Methods. 2021;53(6):2544-2557. [FREE Full text] [doi: 10.3758/s13428-021-01597-3] [Medline: 33954913]
51. Toon E, Davey MJ, Hollis SL, Nixon GM, Horne RSC, Biggs SN. Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents. J Clin Sleep Med. 2016;12(3):343-350. [FREE Full text] [doi: 10.5664/jcsm.5580] [Medline: 26446248]
52. Ena A, Rodríguez A, Woelfe T, Lorscheider J, Granziera C, Kappos L, et al. Elevating the bar by demonstrating cross-device reliability to advance future-proof and hardware-agnostic digital endpoints in multiple sclerosis. Mult Scler. 2024;30(3_suppl):1241-1242. [doi: 10.1177/13524585241269217]
53. Our commitment to transparency of clinical study information. Roche. URL: https://go.roche.com/data_sharing [accessed 2025-03-21]

## Abbreviations

**2MWT:** 2-Minute Walk Test
**BYOD:** bring-your-own-device
**IMU:** inertial measurement unit
**IPS DD:** Information Processing Speed Digit-Digit
**IPS:** Information Processing Speed
**MS:** multiple sclerosis
**OS:** operating system
**PT:** Pinching Test
**SBT:** Static Balance Test
**UTT:** U-Turn Test

*Edited by A Schwartz; submitted 10.06.24; peer-reviewed by B Pueo, U Sinha; comments to author 20.09.24; revised version received 27.11.24; accepted 19.02.25; published 03.04.25*

*Please cite as:*
*Kriara L, Dondelinger F, Capezzuto L, Bernasconi C, Lipsmeier F, Galati A, Lindemann M*
*Investigating Measurement Equivalence of Smartphone Sensor–Based Assessments: Remote, Digital, Bring-Your-Own-Device Study*
*J Med Internet Res 2025;27:e63090*
*URL: https://www.jmir.org/2025/1/e63090*
*doi: 10.2196/63090*
*PMID:*