Formal Assessment of Agreement and Similarity between an Open-Source and a Reference Industrial Device, with an Application to a Low-Cost pH Logger

Evmorfia P. Bataka , Persefoni Maletsika , Christos T. Nakas *

Posted Date: 19 December 2023

doi: 10.20944/preprints202312.1372.v1

Keywords: open-source logger; open-source software; agreement; similarity; pH

Article

Formal Assessment of Agreement and Similarity between an Open-Source and a Reference Industrial Device, with an Application to a Low-Cost pH Logger

Evmorfia P. Bataka [1], Persefoni Maletsika [2] and Christos T. Nakas [1,3,*]

[1] Laboratory of Biometry, Department of Agriculture, Crop Production and Rural Environment, University of Thessaly, Fytokou St., 384 46 Volos, Greece; bataka@uth.gr
[2] Laboratory of Pomology, Department of Agriculture, Crop Production and Rural Environment, University of Thessaly, Fytokou St., 384 46 Volos, Greece; pmalets@uth.gr
[3] University Institute of Clinical Chemistry, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland; cnakas@uth.gr
* Correspondence: cnakas@uth.gr;

Abstract: Open-source devices are nowadays used in a vast number of research fields like medicine, education, agriculture, and sports among others. In this work, an open-source, portable, low cost pH logger, appropriate for in situ measurements was designed and developed to assist in experiments on agricultural produce manufacturing. The device was calibrated manually using pH buffers for values of 4.01 and 7.01. Then it was tested by manually measuring pH from the juice of citrus fruits. A waterproof temperature sensor was added to the device for temperature compensation when measuring pH. A formal method comparison process between the open-source device and a Hanna HI9024 Waterproof pH Meter was designed to assess their agreement. We derived indices of agreement and graphical assessment tools using mixed-effects models. Advantages and disadvantages for interpreting agreement through the proposed procedure are discussed. In our illustration, indices reported mediocre agreement and the subsequent similarity analysis revealed a fixed bias of 0.22 pH units. After recalibration, agreement between the devices improved to excellent levels. The process can be followed in general to avoid misleading or over-simplistic results of studies reporting solely correlation coefficients for formal comparison purposes.

Keywords: open-source logger; open-source software; agreement; similarity; pH

1. Introduction

Open-source devices are becoming very popular and even essential in an increased number of fields, such as education [1], agriculture [2,3] medicine [4], biology [5] among others. The Maker Movement [6] that unfolded after the resurgence of the participatory Web 2.0 [7], the interfusion of Open Source, the decreased cost of electronic parts and other social influences are a few of the examples that contributed to the flourishing of prototype development. Furthermore, this phenomenon was boosted after the launch of development boards like Arduino [8,9] and Raspberry-pi [5] that tend to simplify intricate electronic assemblies by using basic software and programming. Openly accessible tutorials simplify the technical parts, provide visual aids for the wiring, issue the code in each case, and allow users with basic or no experience in electronics and coding, to replicate or customize projects according to their needs. Open-source software and hardware solutions can be used in research and industry. Low acquisition cost and easy customization are two of the most important advantages of using such devices compared to industrial.

Oellermann et al. [10] highlighted three points where open-source electronics aided the scientific community. First, open-source devices help individual researchers by increasing the customization, the efficiency and scalability of the experiments, while increasing data quantity and improving data quality. Second, they assist institutes since the open access to customizable high-end technologies increased the interdisciplinary collaborative networks potential. Third, they succor the scientific community by improving transparency and reproducibility. Also, they help detach research capacity

 doi:10.20944/preprints202312.1372.v1

from funding and escalate innovation. Most of the labs worldwide do not have access to vital funding to keep up with the state-of-the-art lab equipment. Open-source devices contribute to the rapid supply of equipment for the labs with low cost and high level of customizability.

Quality assessment of an in-house built open-source device must be validated by comparing measurements to these of a reference device such as a validated industrial equipment of a reference standard device, in terms of measurement agreement [11]. The comparison is not limited to open-source compared to industrial devices but extends to the comparison of methods in general. "Agreement" measures the "closeness" between readings. Thus, agreement is a comprehensive term that contains both accuracy and precision. Typically, one of the devices/methods is treated as the reference, then agreement concerns a method or measurement comparison study (MCS) of the tested device versus the reference one.

This article presents the development of an open-source device that measures pH of citrus fruit juice and describes the analytical procedure for a method comparison study between the open-source device and its corresponding industrial. The acidity of fleshy fruit, as measured by titratable acidity and/or pH, is an important component of fruit organoleptic quality. Fruit acidity is associated with the presence of organic acids, with malic and citric acids to be the most abundant in most ripe fruits [12]. There is an interrelated relationship between pH and titratable acidity. Titratable acidity is determined by neutralizing the acid present in a known quantity of food sample using a standard base, while the endpoint for titration is usually a target pH (or the color change of a pH-sensitive dye). The titratable acidity of fruits is used, along with sugar content (sweetness), as an indicator of maturity [13]. Citrus is one of the most important commercial fruit crops in the world, and fruit weight, size, acidity and maturity index, harvest time, chemical and nutritional composition are important quality traits for fresh citrus consumption and acceptance by the citrus industry.

Organic acids and sugars expressed as total soluble solids (TSS) vary according to species, varieties, and environmental and horticultural conditions such as climate, rootstock, and irrigation [14]. In citrus fruits, the perceived flavor depends on the combination of taste and aroma, in which the sweet and sour taste attributes are principally result from the presence of sugars and acids in the juice, and the aroma depends on many volatile organic compounds [15,16]. The TSS/acidity ratio has been used worldwide as the main commercial maturity indicator of citrus fruit internal quality. In general, a TSS/acidity ratio of at least 6 or higher is acceptable for commercial marketability, however important differences may exist depending on the citrus species. Ratios acceptable for commercialization usually range from 7-9:1 for oranges and mandarins to 5-7:1 for grapefruits. For lemon this index is not applied, even though in India, the ripening period, juice content and a total acidity of 9% are considered as good indicators of a mature fruit [14].

The benefits of using an open-source device interchangeably with a corresponding industrial are mostly based on the lower cost and configurability of the former. Thus, a method comparison study between the constructed open-source device and a reference industrial one was designed and their agreement and similarity were formally assessed. Measurement ranges where the difference between the two devices are accepatable are discussed. Using recalibration methods the agreement increases.

The paper is organized as follows. The design of the device and its key components are introduced in Section 2.1. Section 2.2. presents the reference device. Section 2.3 describes the five steps to implement a formal statistical method comparison study. The application is discussed in Section 3 and follows the format of Section 2. We end with a discussion and a conclusion. Methodological and further details are given in the Appendix.

2. Materials and Methods

2.1. Design of the device

2.1.1. Hardware

The open-source logger is equipped with two sensors. A pH sensor [17] from Seeed studio and a temperature sensor DS18B20 [18]. A 16-bit analog to digital converter is added to the design to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

improve the precision of the voltage reading since the output of the sensor is analog. The development board for this device is the Adafruit feather proto 32u4. An Adafruit Featherwing logger [19] was added for instant capture of the measurement in a microSD card embedded with a timestamp by pressing a push button. A Nokia 5110 LCD monitor [20] was added to display the values of voltage, pH, temperature, and battery. A 1200 mAH LiPo battery is the main power source of the device which can be charged via micro-USB to USB-A while the device is operating.

The pH Sensor

SEN0169 (Figure 1) is an analog pH meter, specifically designed for Arduino and Arduino-compatible microcontrollers. The electrode is considered as industrial. The sensor has long life (>0.5 years in long-term monitoring mode), is highly accurate (0.1 pH in room temperature), it has fast response ($\leq$ 1 minute), has a measuring range from 0 to 14 pH and includes a gain adjustment potentiometer. The output voltage of the electrode is linear and is capable of long-term monitoring. The sensor has an industrial built and uses a BNC connector and PH2.0 sensor interface. Table A1 (Appendix A) summarizes the technical specifications of the probe. The communication between the sensor and the Microcontroller Unit (MCU) is one-way since the sensor transmits data using an analog MCU pin. Since the 32u4 MCU uses a 10-bit analogue to digital converter (ADC), an ADC1115 16bit ADC and gain amplifier is added to increase the sensor's precision.

Figure 1. DFRobot PH meter (SEN0169) (source: DFRobot official website).

The temperature sensor

DS18B20 is a waterproof digital temperature sensor designed for Arduino or Arduino compatible microcontrollers. According to the manufacturer, since the sensor's signal is digital no signal degradation is present even if the distances between the MCU and the sensor are very long. The sensor provides 9-to-12-bit resolution temperature readings (configurable via software). The communication protocol between the MCU and the sensor is 1-Wire. Multiple DS18B20 sensors can connect on the same 1-Wire bus since they are produced with a unique silicon serial number. Table A2 (Appendix A) summarizes the technical specifications of the sensor.

ADS1115 16bit ADC with gain amplifier

This module is a precision (ADC) with 16 bits of resolution. The first 15 bits are used for the value and the last one for the sign of the value. It is equipped with a voltage reference and an oscillator. It uses I2C communication protocol to interact with the MCU. Four different slave addresses can be selected allowing four different ADS1115 [21] modules to be connected in the same bus. Its operating voltage ranges from 2 to 5.5 Volts. Furthermore, it can converge signals at rates up to 860 samples per second. Its second functionality includes programmable gain amplifier that provides input ranges from inputs to as low as ± 256 mV with increments of 0.0078125 mV, thus measuring both small and large signals with high resolution. Moreover, it offers an input multiplexer which provides two differential or four single-ended inputs. Last, the module operates in continuous conversion mode or a single-shot mode. This means that it automatically powers down in single-shot mode, reducing the power consumption during the measuring periods. To avoid damaging the module, the gain should be set more than or equal to the input voltage of the channel.

 doi:10.20944/preprints202312.1372.v1

2.1.2. Software

DFRobot provides a library for the SEN0169 via GitHub [22] . The library includes a calibration mode. However, the calibration was performed manually due to the MCU's incompatibility. Furthermore, the code was developed without using the library.

The code functionality is described as follows. First, the MCU reads the signal of the pH sensor via the ADS1115, in continuous mode using a single input channel. Second, in case an instantaneous measurement needs to be taken and stored the user will press the push button and the measurement embedded with a time stamp will be stored in the microSD card. The function *button()* and *store()* provide these functionalities. After the calibration procedure, the equation is stored in the sketch and the *measure()* function returns the pH measurement after inserting the input voltage. The function *measure()* returns the proper calibration line, depending on the temperature of the liquid. The sketch is available on GitHub [23].

2.1.3. Calibration Method

The calibration procedure was performed using two pH buffers. Eight measurements were taken. The first two were taken from 4.01 and 7.01 pH buffers when the liquid's temperature was 7.5°C. The same procedure followed for temperatures of 12.5°C, 17.5°C and 22.5°C. The probe was removed from the solution 1 minute after its insertion to reach the response time according to the sensor's datasheet. The temperature of the buffer solution was measured using the DS18B20 temperature sensor. Table A3 summarizes the voltage and their corresponding pH values. Each temperature interval is using the calibration equation of the corresponding midpoint temperatures. For example, equation (1) will be used for the range between 5°C and 10°C. Figure A1 presents the four calibration lines per temperature range.

$$y_{7.5°C} = -6.27615x + 16.4456$$

$$y_{12.5°C} = -6.1349695x + 16.25767$$

$$y_{17.5°C} = -6.0241x + 16.09036$$

$$y_{22.5°C} = -6.04351x + 16.1196615$$

The equations were added to the Arduino sketch in the *measure()* function. PH was automatically calculated depending on the temperature measurement.

2.1.4. Cost of the device

Table A4 summarizes the cost per component and the total cost. The cost can be reduced if parts like the development board and the ADS1115 can be replaced by cheaper equivalents from other brands. Furthermore, the final product is usually not using development boards, removing the inessential parts. Thus, the cost and the device's footprint reduce, especially when the PCB is designed and printed with Surface Mounted Discrete (SMD) electronic parts replacing the through hole equivalents.

2.2. The reference device

The reference device is a Hanna Instruments HI9024 Waterproof pHMeter [24] (Figure 2b). It is a heavy-duty pH meter designed for laboratory use and its accuracy is sustained even under harsh industrial conditions. It can easily be calibrated and has three memorized buffer values (4.01, 7.01 and 10.01). The device has automatic buffer recognition thus avoiding errors during the calibration procedure. Moreover, it is equipped with a temperature compensation function. The temperature can be measured using a temperature sensor probe or can be entered manually. Since there was no temperature sensor available, the temperature was set manually using the DS18B20 sensor which was embedded in the open-source logger. Thanks to its waterproof cylindrical case the temperature

 doi:10.20944/preprints202312.1372.v1

sensor was inserted in the solution that was intended to be measured during the experiment. The specific pH meter model is not available in the market since it is considered obsolete. An equivalent but contemporary model is HANNA HI 99171. Its late 2023 cost in local vendors is around € 585 including shipping costs.

Figure 2. (a) The hand squeezing procedure. (b) Hanna pH meter. (c) DFRobot pH meter probe.

2.3. Designing a Method Comparison Study

To evaluate the open-source device validity, its measurements need to be formally compared with a reference. In other words, a method comparison study needs to be designed to assess the novel device's agreement with the reference device. Five steps can be defined for such studies:

1. Establishment of the experimental design.
2. Exploratory analysis.
3. Assessment of the agreement and similarity between the two devices.
4. Identification of possible sources of disagreement using similarity and repeatability assessment for each device.
5. Recalibration of the novel device to improve the agreement.

2.3.1. Experimental Design

Proper experimental design is of utmost importance for valid results and adequate reproducibility. Repeated (towards intra-variability estimation) and replicate (towards inter-variability estimation) measurements are both multiple response measurements taken at the same combination of factor settings. However, repeated measurements are taken during the same experimental run or consecutive runs, while replicate measurements are taken during identical-conditioned but different experimental runs, which are often randomized. Their differences affect the structure of the dataset and the statistical analysis applied to process the data. In many situations, researchers mistakenly take for granted the sample's independence even though they sample from the same subject. This occurs when the experimental unit is not defined properly and instead of replicates the researchers provide repeated measurements (pseudo-replications).

There are two possible categories of repeated measurements that the present experiment's data fall into. Unlinked and linked data. Following Carstensen et al. [25], unlinked data refer to repeated measurements that are not paired in the sense that the measurements of the two methods are obtained separately. Thus, unlinked data are not necessarily measured concurrently. There is no need for the methods to have the same number of repeated measurements. However, linked data, in which each subject may experience consecutive measurements over time, are paired. Unlike the unlinked data the devices/methods need to have an equal number of paired repeated measurements per one subject but may vary between different subjects. The true value does not need to stay constant over time but there is no systematic effect of time on the paired trajectories beyond the dependence induced in them by the common measurement time.

 doi:10.20944/preprints202312.1372.v1

A well-designed experiment must include a proper definition of the experimental design, the type and number of repeated measurements, the sample size calculation/consideration, and a list of possible covariates. Described methods include covariate information handling.

2.3.2. Exploratory Analysis

A Bland-Altman plot [26] is typically used to assess the data for heteroscedasticity, dependency of the difference from the measurement range, outliers and a linear trend which indicates a correlation between differences and averages. Moreover, a scatterplot can be used as a supplementary plot to investigate the relationship between the two methods. Furthermore, a trellis plot is useful for the visualization of the spread of the repeated values and possible biases of the two methods. A trellis plot [11] is constructed by using the x-axis as the measurement range and the y-axis as the subjects' id. The two methods are differentiated by using two different symbols for each measurement per subject. Interaction plots between subjects and methods (devices), and subjects and time are useful for the researcher to assess graphically the category of repeated measurement. In case there is significant subject x method interaction then an extra term should be added during the modeling process. In case there is significant subject x time interaction then there is a possibility that the data are linked. This can be verified formally using criteria such as AIC, BIC and log-likelihood to assess the model quality.

2.3.3. Statistical tools to assess agreement and similarity.

Mixed-effects and measurement-error models can be fit to the data and their estimated coefficients and variance components are used to produce agreement and similarity indices. These methods go beyond the assessment via standard correlation coefficients given the capacity of handling repeated measures and covariate information. Furthermore, correlation does not imply agreement which is the cornerstone in method comparison studies [11]. Mixed-effects models are a special case of measurement-error models, specifically, if there is evidence in the exploratory analysis that the proportional bias significantly deviates from 1, measurement-error models must be used instead of their mixed-effects counterparts.

The extended Bland-Altman plot can be used during the exploratory analysis step to assess this assumption. If there is a linear trend, then there is evidence of violation of the equal proportional bias assumption of the mixed-effects model. However, this trend might be due to different precisions of the two methods. In any case, the extended Bland-Altman plot be can evaluated using *bland_altman_plot()* function from "*MethodCompare*" [27] package.

The methodology to fit mixed-effects models to the data is described in Appendix B.1.1. which also covers cases when the data are heteroscedastic and when covariates are added. All the steps to prepare the data and implement the models along with their diagnostics are available in an in-house built R-script [23] which is based on [28].

The methodology to fit measurement-error models to the data is described in Appendix B.1.2. which also covers cases when the data are heteroscedastic but does not include covariates. The R-package "*MethodCompare*" [27] can be used to implement the relevant methodology [29–31]. The data must be in wide format. The output includes a list with the estimated bias (differential (fixed) and proportional) including 95% confidence intervals. Moreover, a list of models along with various variables needed for the estimation is returned.

Indices and methods to assess agreement and similarity.

Indices can quantify the agreement and similarity between two or more devices. There are two categories of indices. The absolute (or unscaled) and the relative (or scaled) indices [32]. A detailed review about agreement indices can be found in [33].

Absolute indices report measures according to the magnitudes of the actual data. They are unscaled and independent of between-sample variation.

The total deviation index (TDI) is used here for the evaluation of the agreement and similarity between two method/devices (inter-agreement). Specifically, TDI is an index that captures a

 doi:10.20944/preprints202312.1372.v1

predefined proportion (p) of data within a boundary (δ) from target values, defined by $TDI(p) < \delta$.

Two measurement methods may be considered to have sufficient agreement if a large proportion of their differences is small. Thus, we define p as the proportion of their differences and δ as the sufficient difference. Its estimate can be evaluated using (B7).

TDI can be also used for the evaluation of the intra-agreement for each device separately. The estimates can be evaluated using (B14, B18).

Relative indices are scaled values on a predefined range and usually lie between -1 and 1. The concordance correlation coefficient (CCC) is the most popular index for assessing agreement between quantitative measurements (inter-agreement). There is perfect agreement when CCC = 1, no agreement when CCC = 0 and perfect negative agreement when CCC = -1. Its estimate can be evaluated using formulas (B8, B9). CCC can be also used for the evaluation of the intra-agreement for each device separately. The estimates can be evaluated using formulas (B15, B19).

Moreover, the 95% limits of agreement produce an interval within which 95% of differences between measurements made by the to methods/devices are expected to lie [26].

An in-house built R-script [23] implementing the relevant methods [28] may be used to evaluate CCC and TDI along with their corresponding bounds. Moreover, TDI evaluation and its upper bound, based on an alternative formulation of mixed-effects models [34] can be implemented [35]. CCC evaluation and confidence intervals for inference, instead of a lower bound, using an alternative formulation of mixed-effects models [36,37] can be implemented using the "*cccrm"* package [38]. The limits of agreement can be evaluated and presented graphicaly by the *"methodCompare"* package via the *bland_altman_plot()* function along with the corresponding extended Bland-Altman plot. A wide data format to implement *bland_altman_plot()* is needed. The package "*blandr"* [39] can be used to evaluate the limits of agreement along with their confidence intervals, superimposed on a Bland-Altman plot.

Moreover, the bias plot (*bias_plot()* function [27]) from the "*methodCompare"* package, evaluates the differential (fixed) and proportional bias (described in Appendix B.1.2.) and offers a useful display that quantifies systematic bias (fixed and proportional) along the measurement range.

2.3.4. Investigating possible sources of disagreement

Assessing Similarity

Early research assessing similarity measures was focused on paired data [40]. Precision and accuracy via the quantification of fixed and proportional bias, along with the precision ratio were proposed as measures of similarity [41].

For mixed-effect models, only the fixed bias can be evaluated since proportional bias is assumed to be equal to 1. To implement the standard methodology to evaluate the fixed bias and precision ratio [11] an R-script is available online [23]. The formal methodology for similarity assessment can be found in Appendix B.2.2.

For measurement-error models, similarity can be evaluated using a bias plot (discussed in Section 2.3.3) and a precision plot [29–31]. The precision plot can be implemented using the "*methodCompare"* [27] package via the *precision_plot()* function.

Assessing Repeatability

The evaluation of repeatability is essential and can be used to identify possible sources of disagreement. It is considered as intra-method agreement and is an essentialpart of the agreement study. When a method/device has low intra-method agreement it will most probably have low inter method agreement suggesting poor overall agreement of methods or devices.

For mixed-effect models, CCC, TDI and corresponding 95% limits of agreement can be used to assess intra-method agreement. These are evaluated for each method/device separately and assess the agreement between repeated measurements with the same device. Implementation of relevant methods [11,28] is possible using an online R-script developed by the first author [23].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

For measurement-error models, repeatability can be assessed graphically via a bias plot (discussed in Section 2.3.3) by investigating the spread of the measurements of each method/device. Repeatability can also be assessed using a Trellis-plot.

2.3.5. Recalibration Methods

For the mixed-effects model, a recalibration procedure is performed by subtracting the fixed-bias. The relevant methodology [28] can be implemented using the R-script available online [23].

For the measurement-error model, a recalibration procedure is described in [29–31] and the function *compare_plot()* from the "*MethodCompare*" package can be used to implement it. Appendix B4 describes the procedure.

Figure 3 summarizes the workflow of a method comparison study.

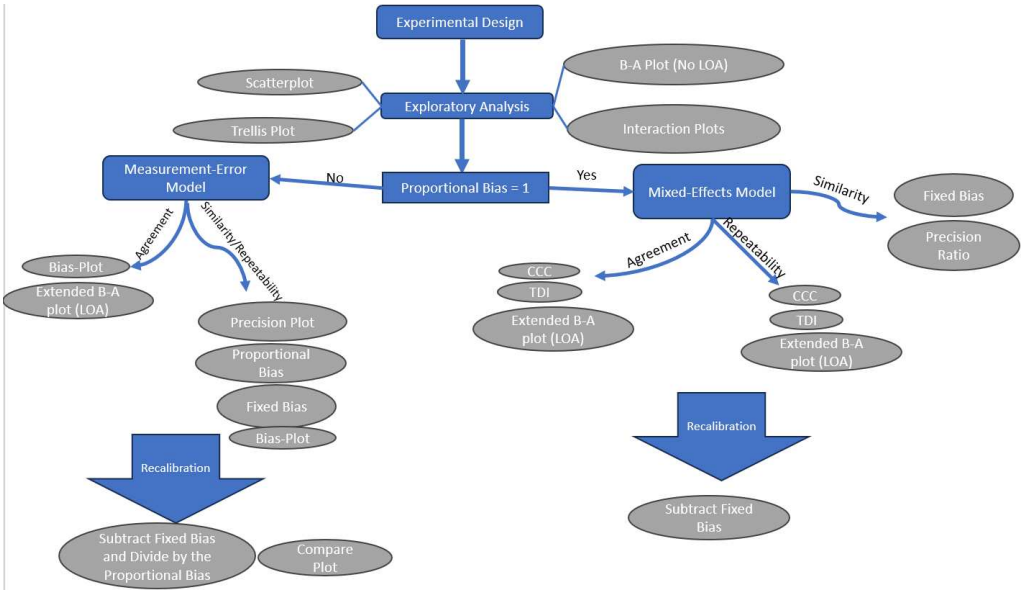


Figure 3. The workflow of a method comparison study.

3. Application

3.1. Experimental Design

The solution (juice) was extracted from two varieties of citrus fruits. Each fruit is considered as an experimental unit. In total, 15 grapefruits and 15 juice oranges (Valencia variety) were used. Each unit was hand-squeezed (Figure 2a), and its juice was measured by the open-source device and by a Hanna HI 9024 pH meter (which defined the reference device). The order of measurements was randomized using R's *sample* function, and 10 repeated measurements for each fruit were collected by a single reader/operator (EB). Repeated measurements were sequentially taken. First, the *sample* function was used to define the instrument that will measure first. The other instrument was used next. Nine more measurements of the same juice were taken by first cleaning each instrument using deionized water and then taking the measurement.

The data are considered linked since they are paired over the measurement times. Figure 2c displays the open-source pH sensor and the measurement procedure. Table 1 summarizes the experimental design information. The type of the fruit (grapefruit or juice orange), temperature, quantity of the juice, and the instruments' sequence were considered as covariates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Table 1. Summary of the experimental design information. The sample size is 30, since each fruit is considered as one experimental unit. There were ten repeated measurements in a balanced design. The data are considered as linked.

Experimental Design	
Experimental Unit	Fruit
Repeated Measurements	Yes, sequentially
Number of repeated measurements	10
Data category	Linked data
Sample size	30
Balanced/Unbalanced measurements	Balanced
Possible Covariates	Temperature, juice quantity, instrument turn

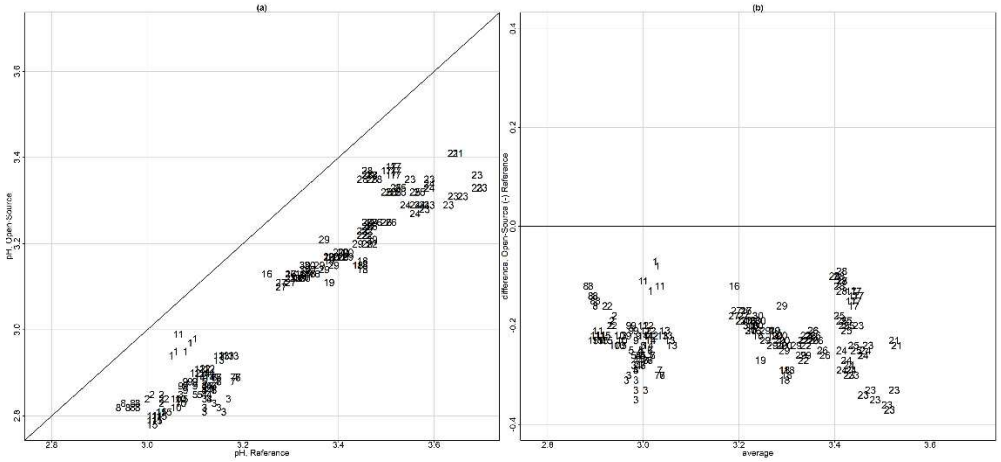
3.2. Exploratory Analysis

Exploratory data analysis involved three different depiction approaches. Figure 4a displays a scatterplot for the pH measurements of the reference versus the open-source device. To avoid using the same plotting symbol per subject and visualize the repeated measurements, each subject is represented by a unique id number and the repeated values share the same id subject symbol. Using this approach, a dependence structure is depicted. A systematic underestimation of the open-source device for pH measurements is apparent. There are two clusters formed in the data. The lower left corresponds to the grapefruit pH while the upper right to the orange juice.

Figure 4b displays a Bland-Altman plot (averages vs differences) without the limits of agreement. For higher values of pH, the differences seem to have slightly higher spread compared to lower values of pH. This is a sign of possible heteroscedastic errors. There is no obvious trend in the Bland-Altman plot suggesting a common scale for the assays, verifying the common scale assumption for the mixed-effects model. This is also obvious in the extended Bland-Altman plot, which was produced using the "MethodCompare" package (Figure 6).

Figure 5 displays a trellis plot. The vertical axis is divided into rows and each row displays all the repeated measurements for one subject and both devices using method-specific colors. Blue color represents the measurements for the reference device while yellow represents the measurements for the open-source device.

Since the repeated measurements are plotted in one row, within-subject variability is visible and easy to compare with the between-subject variability. The open-source device shows slightly less within-subject variation compared to the reference. The between-subject variation ranges between 2.78 to 3.7 and a summary is presented in Table 2. A consistent bias is also visible in the graph, suggesting a constant fixed bias throughout the measurement range. The open-source device underestimates the pH measurements by approximately 0.22 units.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Figure 4. (a) Scatterplot for the pH measurements of the reference versus the open-source device. Each subject is represented by a unique id number and the repeated values share the same id subject number symbol. A systematic underestimation of the open-source device for pH measurements is apparent. There are two clusters formed in the data. The lower left corresponds to the grapefruit pH while the upper right corresponds to the juice oranges. (b) Bland-Altman plot (averages vs differences) without the limits of agreement. For higher values of pH the differences seem to have slightly higher spread compared to lower values of pH. This is a sign of possible heteroscedastic errors. There is no obvious trend suggesting a common scale for the assays, verifying the common scale assumption for the mixed effects model.

Figures A2a and A2b display the interaction plots for subject x method and subject x time respectively. For the subject x method interaction plot the average per subject for every measurement is plotted on the vertical axis and each method on the horizontal (Figure A2a). There is evidence of a significant subject x method interaction since the lines intersect. Figure A2b displays the subject x time interaction. The repeated measurements are averaged per method for each subject (vertical axis) and the time points are displayed on the horizontal axis. Some lines intersect, providing evidence of possible interaction between subjects and time.

Table 2. The minimum, 1st quartile, median, mean 3rd quartile, and maximum pH values per device. The measurement range lies between 2.78 and 3.7.

	Minimum	Q1	Median	Mean	Q3	Max
Open-Source	2.780	2.868	3.045	3.053	3.220	3.410
Reference	2.040	3.110	3.220	3.275	3.460	3.700

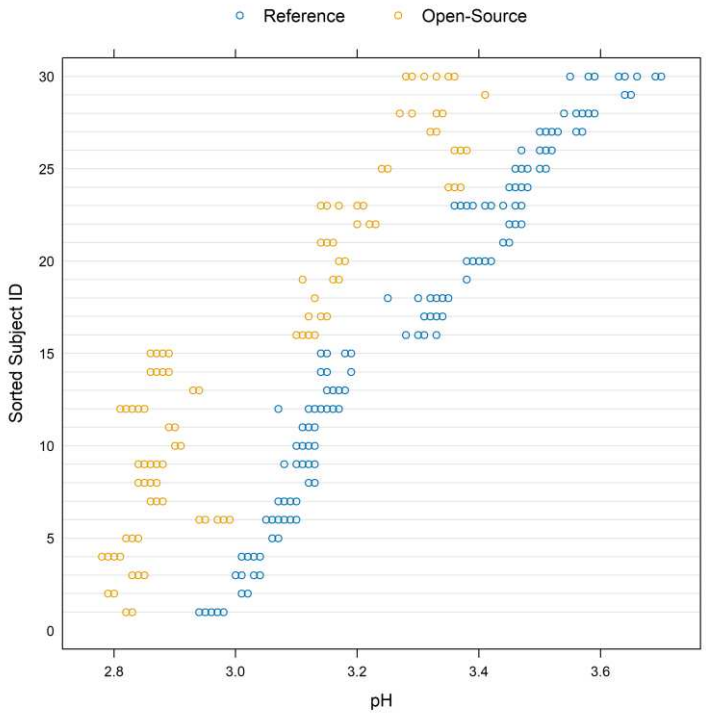


Figure 5. Trellis plot for the pH measurements. The vertical axis is divided into rows and each row displays all the repeated measurements for one subject using method-specific colors. Blue color represents the measurements for the reference while yellow represents the measurements for the open-source device. The open-source device underestimates pH measurements for most of the subjects by approximately 0.22 units. Moreover, the within subject-variation for the open-source device is less than the reference's. The between-subject variability ranges between 2.78 and 3.7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

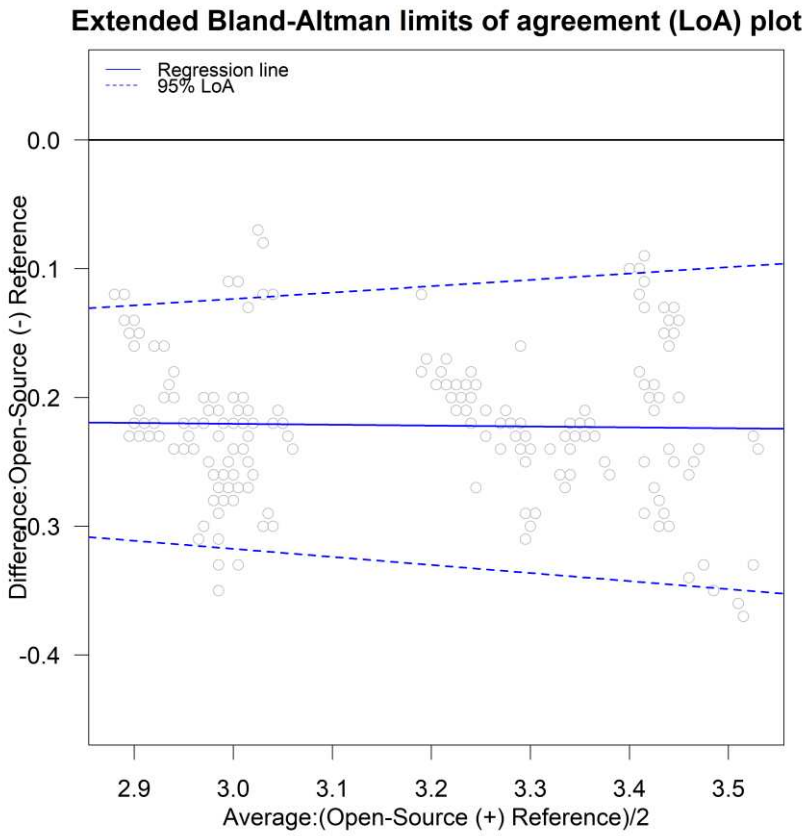


Figure 6. Extended Bland-Altman plot and LOA. There is slight evidence of heteroskedastic errors. No trend is apparent; thus, a common scale is assumed for the assays.

3.3. Statistical tools to assess agreement and similarity.

Initially, the data were fit with the homoscedastic model with no covariates for linked data (B1), then the corresponding heteroscedastic. However, the additional computational burden provided by the subject to occasion interaction for the linked data hindered the procedure to calculate the confidence bounds for the indices. Thus, the unlinked homoscedastic and heteroscedastic models were chosen to proceed with the analysis. There is no obvious sign of a fan shape. AIC and BIC were subsequently calculated, and the heteroscedastic model was chosen (Table 3).

At a subsequent stage, model (B4) was used to fit the data, which includes the covariates without interactions. According to the AIC and BIC criteria, the model without covariates was preferred. Table 4 displays AIC, BIC, Log-Likelihood and degrees of freedom for the heteroscedastic models with and without covariates.

To account for heteroscedasticity, a sequence of 20 values starting from 2.78, which is the minimum value for the average values of the two methods, and 3.7 which is the maximum value for the average of the two methods was created. Then, the variance function was defined as $g(u_i, \delta) = |u_i|^{\delta}$, where $\tilde{u}_i = h(\bar{y}_{i1}, \bar{y}_{i2}) = \frac{\bar{y}_{i1.}+\bar{y}_{i2.}}{2}$. The variance function parameter $\tilde{u}_i$ can also be chosen as the average values per subject of the reference device. No significant changes are reported regardless of the choice of $\tilde{u}_i$. The parameter $\delta_1 = 4.07$ for the reference and $\delta_1 = 3.28$ for the open-source device. The model's counterparts are displayed in Table 5.

Diagnostics for the optimal model (Figure A3, Appendix A) display the standardized residuals on the horizontal axis vs the quantiles of the standard normal distribution. The plot reveals a slight deviation from the normal distribution. The standard errors for the estimates are reasonable thus the agreement and similarity indices evaluation proceeds using this model.

Substituting the ML estimates from Table 5 in (B3) to obtain the fitted distribution (Y_1, Y_2) given the pH level $\tilde{u}$ yields

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

$$N_2\left(\begin{pmatrix}-0.22\\3.05\end{pmatrix},\begin{pmatrix}0.04062325+2.615606\times10^{-8}\tilde{u}^{4.07} & 0.03934918\\0.03934918 & 0.04062325+7.374885\times10^{-8}\tilde{u}^{3.28}\end{pmatrix}\right),$$

while for D given $\tilde{u}$:

$$D|\tilde{u}\sim N_1(-0.22, 0.002548149+2.615606\times10^{-8}\tilde{u}^{4.07}+7.374885\times10^{-8}\tilde{u}^{3.28}).$$

The intra-method differences distribution given $\tilde{u}$ are produced by substituting the parameters from Table 5 in (B13):

$$D_1|\tilde{u}\sim N_1(0, 5.231212\times10^{-8}\tilde{u}^{4.07}),$$

$$D_2|\tilde{u}\sim N_1(0, 1.474977\times10^{-8}\tilde{u}^{3.28}).$$

D_1 denotes the differences for the reference and D_2 denotes the differences for the open-source device.

Assessment of agreement

Using formula (B6) to calculate the limits of agreement substituting the model's counterparts in Table 5 and the variance function, the inter- and intra- device agreement is displayed in Figure 7. For the inter-agreement of the devices, Table 6 summarizes the ranges for the 95% limits of agreement for pH data as a function of the magnitude of measurements. The inter-method limits, based on the distribution of D, are centered at -0.22. For lower pH values the LOA are narrower compared to the higher pH values and range from $[-0.3464, -0.3237]$ for lower LOA and $[-0.1193, -0.0966]$ for upper LOA. The intervals reveal a systematic underestimation of the pH measurements from the open-source device. Figure 8 illustrates the Bland-Altman plot and limits of agreement along with their corresponding confidence intervals. The "*Blandr*" R package was used to produce the plot.

Table 3. AIC, BIC, Log-Likelihood and degrees of freedom for the homoscedastic and the heteroscedastic model. The heteroscedastic model is slightly better than the homoscedastic model. Either can be chosen. In this case the heteroscedastic is chosen for illustrative purposes. The analysis for the homoscedastic model can be found in Appendix C.

	AIC	BIC	Log-Likelihood	Degrees of Freedom
Homoscedastic	−2945.889	−2919.507	1478.945	6
Heteroscedastic	−2999.128	−2963.952	1507.564	8

Table 4. AIC, BIC, log-likelihood and degrees of freedom for model selection. The model which does not include covariates is slightly better than the one which includes the covariates. Thus, the model with no covariates is selected for further analysis and indices calculation.

Covariates	AIC	BIC	Log-Likelihood	Degrees of Freedom
No	−2999.128	−2963.952	1507.564	8
Yes	−2641.498	−2597.528	1330.749	10

Table 5. Model counterparts. The fixed bias is estimated to be equal to -0.13. Recalibration actions can be taken to improve the agreement by adding the estimated constant. The standard errors are reasonable for the model's validity.

Parameter	Estimate	SE	95% CI
β_0	−0.22	0.01	$[-0.24, -0.20]$
μ_b	3.27	0.04	[3.20,3.35]
$log(\sigma_b^2)$	−3.24	0.26	$[-3.75, -2.72]$
$log(\psi^2)$	−6.67	0.26	$[-7.18, -6.15]$
$log(\sigma_{\varepsilon_1}^2)$	−17.46	1.61	$[-20.61, -14.31]$
$log(\sigma_{\varepsilon_2}^2)$	−16.42	1.61	$[-19.58, -13.27]$
δ_1	4.07	0.70	[2.71,5.44]
δ_2	3.28	0.70	[1.91,4.65]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Table 6. Estimates of 95% limits of agreement for the inter-method agreement for pH measurements as a function of their magnitude. The inter-method limits, based on the distribution of D, are centered at -0.22. For lower pH values the LOAs are narrower comparing to the higher pH values and range from [-0.3464, -0.3237] for the lower LOA and [-0.1193, -0.966] for the upper LOA.

Limits of Agreement (Inter-Method)						
	Minimum	Q1	Median	Mean	Q3	Max
Lower	−0.3464	−0.3369	−0.3305	−0.332	−0.3263	−0.3237
Upper	−0.1193	−0.1167	−0.1125	−0.112	−0.1060	−0.0966

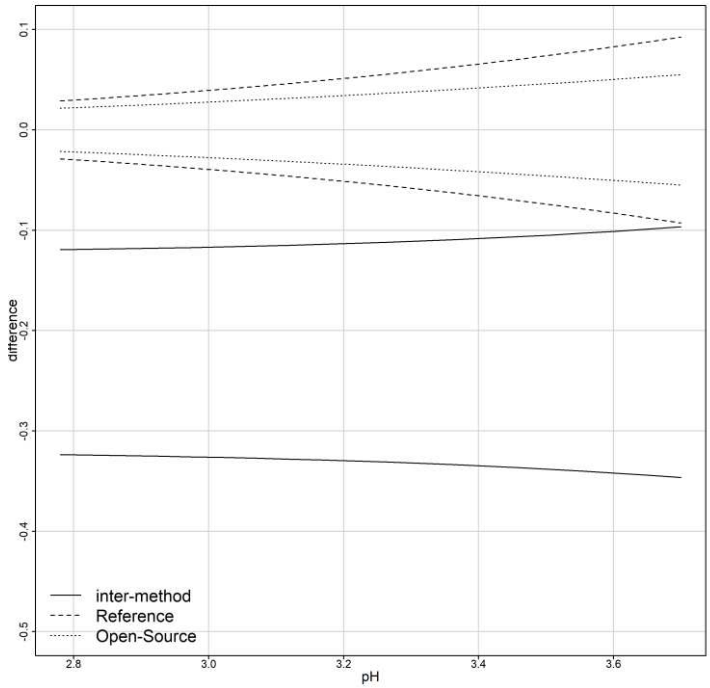


Figure 7. 95% Limits of inter- and intra- method agreement. The intervals reveal a systematic underestimation of the pH measurements from the open-source device. For the 95% limits of the intra-method agreement the open-source device has narrower LOA compared to the reference pH meter; thus, the open-source device has better repeatability. For lower pH values the LOA are narrower compared to the higher pH values.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

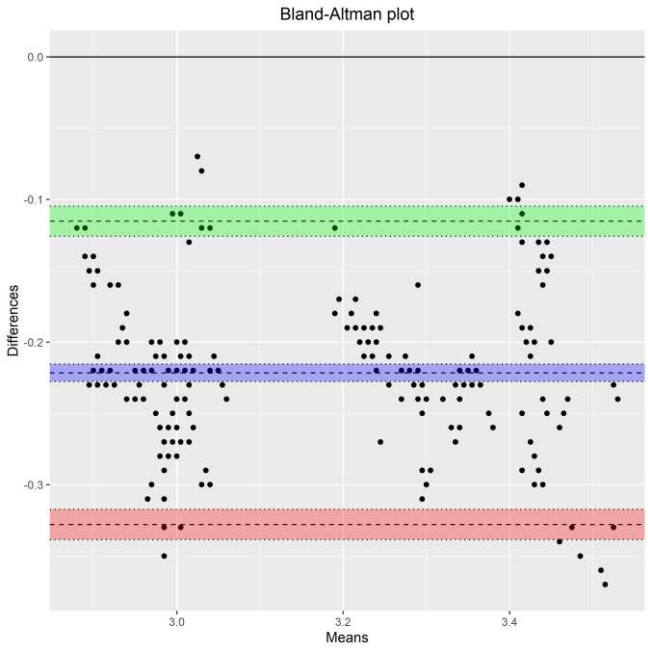


Figure 8. Bland – Altman plot using the "blandr" package in R. Apart from the limits of agreement and the mean difference, their corresponding confidence intervals are plotted.

Table 7 presents CCC and TDI estimates, and lower and upper confidence bounds respectively before and after recalibration. Before recalibration the estimates for CCC range between 0.5970 and 0.6032 while the corresponding lower confidence bounds range between 04776 and 04839 throughout the pH measurement range. TDI (0.9) estimates range between 0.2883 and 0.3031 and their corresponding upper confidence bounds range between 0.3095 and 0.3232 throughout the pH measurement range.

Figure 9a presents one-sided 95% pointwise confidence bands for CCC as a function of the magnitude of the measurements. The solid line represents CCC lower confidence bound for the inter-method agreement and ranges between 0.4776 and 0.4839. CCC lower band decreases as the pH level increases. Thus, the agreement becomes progressively worse but only by a small amount. The inter-method agreement is not considered to be satisfactory. Figure 9b presents the one-sided 95% pointwise upper confidence bands for inter- and intra- method versions of TDI (0.9) and their reflections over the horizontal line at zero. For the inter-method agreement TDI (0.9), which is represented by the solid line, upper confidence bound ranges between 0.3095 and 0.3232. As the pH level increases from 2.78 to 3.7, TDI increases. The bound of 0.3232 shows that 90% of differences in measurements from the devices fall within ± 0.3232 when the true value is 3.7. Such a difference is unacceptably large for many applications. The bounds of 0.3095 and 0.3232 are, in proportional terms, 8.36 and 8.74 % of the true value respectively. A non-significant difference appears for the inter-method agreement throughout the pH measurement range. The similarity evaluation reveals that a difference in the means of the devices is a contributor to disagreement. TDI and CCC improve after recalibration.

Overall, as the magnitude increases TDI increases and CCC decreases. This means that the inter-method agreement becomes worse as the magnitude increases.

Following an alternative approach [35] to calculate TDI (0.9), the estimates are similar with equivalent conclusions before and after recalibration. The same applies for CCC [36].

Figure 10 displays the Bias-plot. The proportional bias is 0.965 (95% CI $[0.9352, 0.9938]$) and the fixed bias is -0.1052 (95% CI $[-0.2013, -0.091]$). The fixed bias estimate is different compared to the standard estimate [11] because the parameter estimation method is different. However, the red solid line which corresponds to the total bias confirms the findings that follow, presented in Table 9, Section 3.5, since the total bias ranges from [-0.235, -0.21].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Table 7. CCC and TDI estimates with their corresponding lower and upper confidence bounds throughout the pH measurement range before and after recalibration. CCC and TDI estimates after recalibration. Both indices improve significantly and report excellent agreement between the devices.

Before Recalibration		
Grid	CCC	Lower Confidence Bound
2.78	0.5970	0.4776
3.7	0.6032	0.4839
Grid	TDI	Upper Confidence Bound
2.78	0.2883	0.3095
3.7	0.3031	0.3232
After Recalibration		
Grid	CCC	Lower Confidence Bound
2.78	0.9943	0.9909
3.7	0.9949	0.9917
Grid	TDI	Upper Confidence Bound
2.78	0.1051	0.1244
3.7	0.1114	0.1295

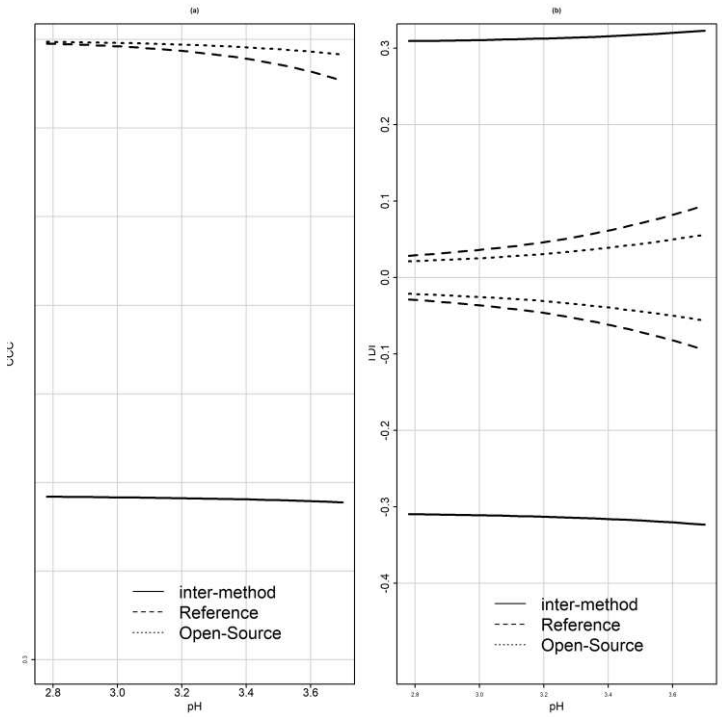


Figure 9. (a) One-sided 95% pointwise confidence bands for CCC as a function of the magnitude of the measurements. The solid line represents CCC lower confidence bound for the inter-method agreement and ranges between 0.4776 and 0.4839. The agreement is considered insufficient. The dashed and dotted line represents the intra-method one-sided 95% pointwise confidence band for the reference and open-source respectively. For the reference the band ranges between 0.9979 and 0.9987 and for the open-source between 0.9986 and 0.9997. The intra-method agreement for both devices is considered excellent. (b) One-sided 95% pointwise upper confidence bands for intra- method versions of TDI (0.9) and their reflections over the horizontal line at zero. The open-source device has higher intra- method agreement.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

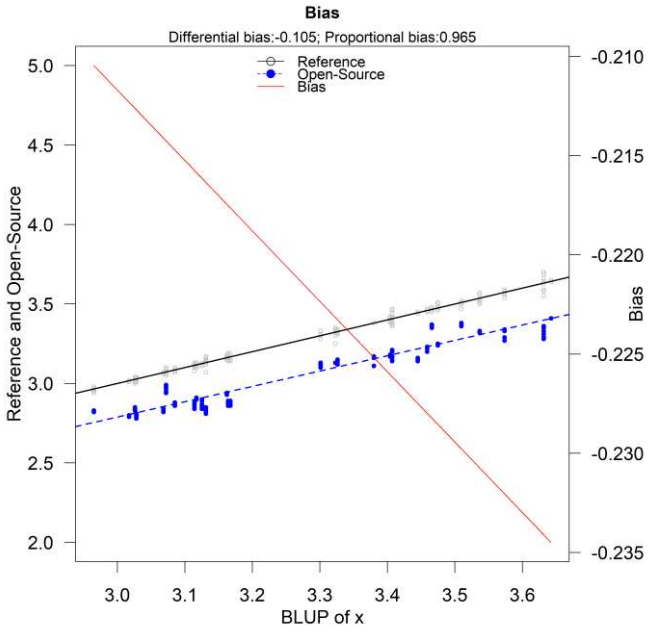


Figure 10. Bias plot of the reference versus the open-source device. The proportional bias is 0.965 (95% CI [0.9352, 0.9938]) and the fixed bias is estimated at -0.1052 (95% CI [-0.2013, -0.091]).

Table 9. Precision ratio estimates as a function of magnitude of measurement. The fixed bias is -0.22 units for the open-source device compared to the reference. The open-source device underestimates the pH measurement by 0.22 units the entire CI is below zero.

Similarity Assessment			
Grid	2.78	Lambda Estimate 1.7987	Confidence Interval [1.03969,2.737423]
	3.7	2.832176	[1.770244,5.471250]
	Fixed Bias	Estimate -0.221497	Confidence Interval $[-0.239736, -0.203257]$

3.4. Investigating possible sources of disagreement

Similarity Assessment

For the assessment of similarity fixed bias and precision ratio are estimated. Fixed bias represents the difference in means of the two devices under the equal scale assumption. Since the errors are heteroscedastic and the precision is defined as the error variance of the reference over the error variance of the open-source, the precision as a function of magnitude of measurement is displayed on Figure 11. Table 9 summarizes the two indices. The open-source device is twice to three times more precise than the reference. The fixed bias is -0.22 units for the open-source device compared to the reference. The open-source device underestimates the pH measurement by 0.22 units since the entire interval is below zero. The open-source device can be considered to be of higher precision. These findings are consistent with the exploratory analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

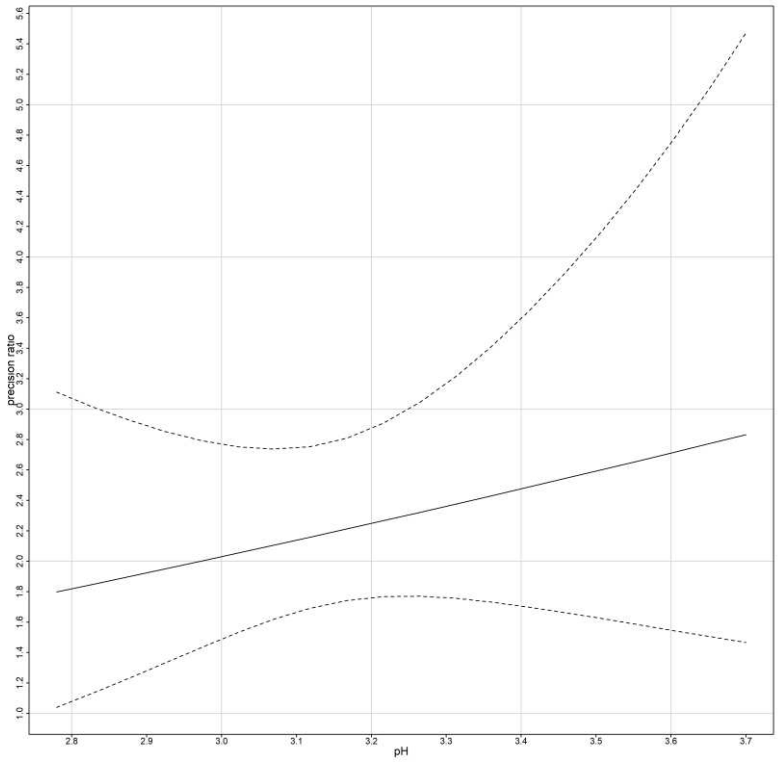


Figure 11. Precision ratio along with corresponding 95% confidence intervals. The open-source device is more precise than the reference, especially as the magnitude of the measurements increases.

Evaluation of Repeatability

CCC, TDI and the limits of agreement are calculated for the intra-agreement of each device separately. Figure 7 displays the limits of agreement as a function of the magnitude of measurement. The limits of agreement for the open-source device (dotted lines) are included in the reference's LOA (dashed lines). Table 10 summarizes the ranges for the 95% limits of Agreement for pH data as a function of magnitude of measurement. The open-source device LOA are narrower comparing to the reference suggesting a better repeatability. Based on the distributions of D_1 and D_2, the intra-method limits are centered at zero. In Figure 9a, the CCC index is presented for inter- and intra-method agreement. The dashed and dotted lines represent the intra-method agreement for the reference and open-source device respectively. For the reference device the upper bound ranges between 0.9979 and 0.9987 and for the open-source between 0.9986 and 0.9997. The intra-method agreement for both devices is considered excellent. However, the open-source device has higher intra-method agreement. This conclusion is expected since the similarity assessment reported smaller error variation for the open-source device.

Figure 9b, illustrates TDI (0.9). For the open-source device, which is represented by the dotted line, TDI (0.9) lower bound ranges between 0.0250 and 0.0504 while for the reference, dashed line, between 0.0479 and 0.0606. The interpretation for TDI (0.9) is as follows: the bound of 0.0250 implies that 90% of the time the difference between two replications of the open-source device on the same subject falls within ± 0.0250 when the true pH value is 2.78. The TDI bounds for both devices are only 0.9 – 1.81 % of the magnitude of measurement, indicating a high degree on intra-method agreement. Table 11 displays CCC and TDI (0.9) along with their corresponding bounds for the minimum and maximum range of the measurements. High intra-method agreement CCC values reflect that the within-subject variations for both assays are very small compared to the between-subject variation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Table 10. 95% limits of agreement for intra-method Agreement for pH data as a function of magnitude of measurement. The open-source device has narrower LOA compared to the reference pH meter. Thus, the open-source device has better repeatability. Based on the distributions of D_1 and D_2, the intra-method limits are centered at zero.

Limits of Agreement (Intra-Method)						
Reference	Minimum	Q1	Median	Mean	Q3	Max
Lower	−0.0926	−0.0713	−0.0540	−0.0564	−0.0400	−0.0289
Upper	0.0289	−0.0400	0.0540	0.0564	0.0713	0.0926
Open-Source	Minimum	Q1	Median	Mean	Q3	Max
Lower	−0.0550	−0.044	−0.0356	−0.0366	−0.0280	−0.0216
Upper	0.0216	0.0280	0.0356	0.0366	0.0446	0.0550

Table 11. Summary for CCC and TDI(0.9) for the intra- method agreement.

Grid	CCC for repeatability: Reference	Lower Confidence Bound
2.78	0.9986	0.9979
3.7	0.9992	0.9987
Grid	**CCC for repeatability: Open-Source**	**Lower Confidence Bound**
2.78	0.9991	0.9986
3.7	0.9998	0.9997
Grid	**TDI for repeatability: Reference**	**Upper Confidence Bound**
2.78	0.0422	0.0479
3.7	0.0543	0.0606
Grid	**TDI for repeatability: Open-Source**	**Upper Confidence Bound**
2.78	0.0221	0.0250
3.7	0.0453	0.0504

3.5. Recalibrating the open-source device

The similarity evaluation reveals that the fixed bias (difference in the means) contributes to the disagreement between the two devices. Recalibration of the open-source devices by subtracting -0.22 from its measurements makes the mean difference zero and improves the extend of agreement substantially. Table 7 reports CCC and TDI estimates and confidence bounds after recalibration. CCC improves significantly. The lower confidence bands range from 0.9909 to 0.9917 revealing excellent agreement throughout the measurement range. TDI also improves and ranges from 0.1244 to 0.1295 throughout the measurement range. The agreement for this case study is considered acceptable. TDI (0.9) and CCC were also calculated after recalibration following Escaramis et al. [26] and Carrasco [36] respectively (Table 8). They are both close to Table 7 values with TDI (0.9) and CCC being lower compared to Table 7 values.

4. Discussion

A portable open-source device that measures the pH of the juice of grapefruits and oranges was designed and constructed for laboratory experiments and in situ measurements. To evaluate its functionality, a method comparison study between the open-source device and a corresponding industrial was designed. The statistical analysis to assess their agreement was based on indices and graphical methods using mixed-effects models. The agreement indices evaluated were the Concordance Correlation Coefficient (CCC) and the Total Deviation Index (TDI). TDI estimates and confidence bounds were evaluated using (B1) and methodology described in Section 2.3 [11]. There were small differences between the two methods probably due to the mixed-effects models' different formulation.

Overall, agreement between the two devices is not satisfactory but improves to excellent levels after recalibration since the main source of disagreement is the fixed bias (0.22 pH units).

Further experiments can be conducted to investigate the agreement for an extended range of measurements and a greater variety of fruits or other applications that include soil pH or substrate

 doi:10.20944/preprints202312.1372.v1

pH in soilless cultivations. An R-Script, schematics and Arduino code for researchers to follow the proposed methodology [42] and develop the open-source device are available [23].

5. Conclusions

This paper highlights the assessment of open-source devices, regarding their functionality and the validity of their measurements. The most effective way to validate the measurements of a novel device is to compare them to established commercial/industrial ones. The official and reliable process to accomplish this task is to design and apply a method comparison study which includes proper experimental design and statistical tools to assess the agreement and similarity between the two devices. This methodology is applied mostly in medical research [43,44] but not limited to it.

Parts of the proposed guide are described in the literature [45], but restricted to the Bland-Altman plot and ICC. The current research proposes a step-by-step procedure to validate open-source devices, including the experimental design, descriptive statistics and a variety of formal statistical assessment and encourages the development of a protocol applied to this highly blooming field.

Author Contributions: "Conceptualization, E.B. and C.N.; methodology, E.B., P.M., and C.N.; software, E.B.; validation, E.B.; formal analysis, E.B.; investigation, E.B., P.M., and C.N.; resources, E.B.,P.M., and C.N.; data curation, E.B.; writing—original draft preparation, E.B.; writing—review and editing, E.B., P.M., and C.N.; visualization, E.B.; supervision, C.N.; project administration, E.B. and C.N.; funding acquisition, E.B. and C.N. All authors have read and agreed to the published version of the manuscript." Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: "This research was funded by the European Social Fund".

Institutional Review Board Statement: "Not applicable."

Informed Consent Statement: "Not applicable."

Data Availability Statement: Data available online: https://github.com/kersee112358/Chapter-5----Ph-logger-sensor

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: "The authors declare no conflict of interest."

Appendix A

Table A1. Technical Specifications of pH sensor SEN0169.

Specifications	
Module Power	3.3 or 5 Volts
Measuring Range	0 − 14 pH
Measuring Temperature	0~60°C
Accuracy	±0.1pH (25°C)
Response Time	≤ 1minute
Interface	Analog Output

Table A2. Waterproof DS18B20 digital temperature sensor specifications.

Specifications	
Module Power	3.3 or 5 Volts
Measuring Range	−55°C to 125°C
Accuracy	±0.5°C from −10°C to 85°C
Resolution	9 to 12 bits ADC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Interface	1-Wire
Steel tube dimensions	6mm diameter by 35mm long

Table A3. Temperature of the pH buffer during the calibration procedure. Eight measurements were taken. Four for 4.01 pH buffer and four for 7.01 pH buffer.

Temparature	Temperature Range (°C)	Voltage	pH
7.5ºC	[5, 10]	1.983	4.01
13.5ºC	(10, 15]	1.505	7.01
17.5ºC	(15, 20]	1.998	4.01
22.4ºC	(20, 25]	1.509	7.01
7.5ºC	[5, 10]	2.007	4.01
13.5ºC	(10, 15]	1.509	7.01
17.5ºC	(15, 20]	2.010	4.01
22.4ºC	(20, 25]	1.514	7.01

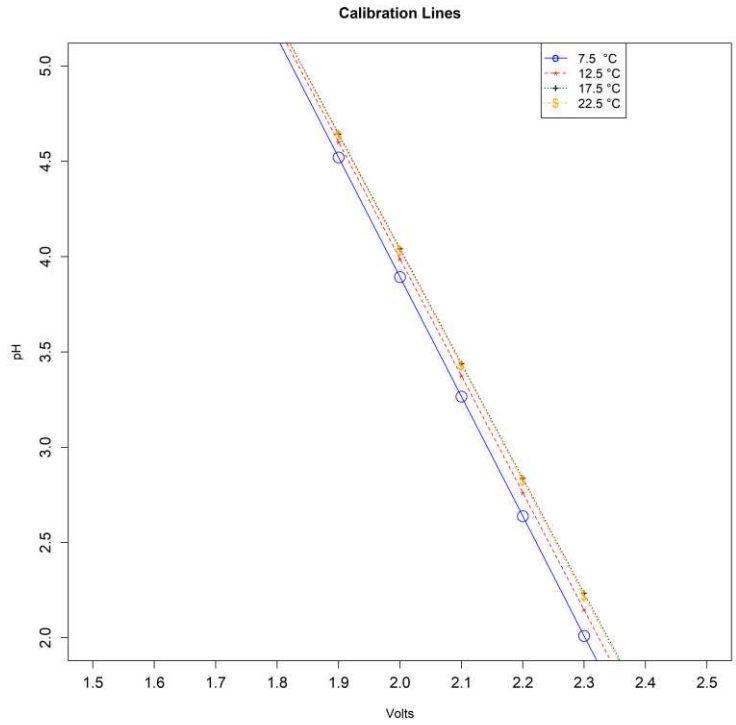


Figure A1. Calibration curves for 7.5ºC, 13.5ºC, 17.5ºC and 22.4ºC.

Table A4. Device cost (€) in late 2023 local vendor prices and international vendor Mouser.

Item	Cost (€)
DFRobot PH meter Pro Kit (SEN0619)	70
Temperature Sensor DS18B20	3
Adafruit Feather proto 32u4	18.75
Adafruit Adalogger FeatherWing	8.41
Adafruit ADS1115 ADC 16bit	18.4
microSD Card	3
Coin Cell Battey	1
IP66 Enclosure Box	5
PCB	1.5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Nokia 5110 module	4.68
Others (Wires, Solder)	2
Total	127.63

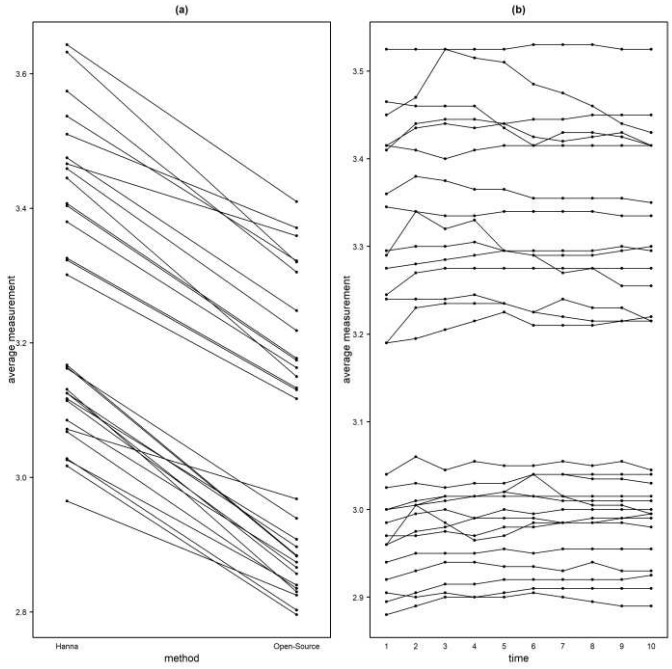


Figure A2. (a) Interaction plot between method and subjects. There is evidence of subject x method interaction since the lines intersect. (b) Interaction plot between subjects and time. A few of the lines intersect, providing evidence of possible, but not strong interaction between subjects and time.

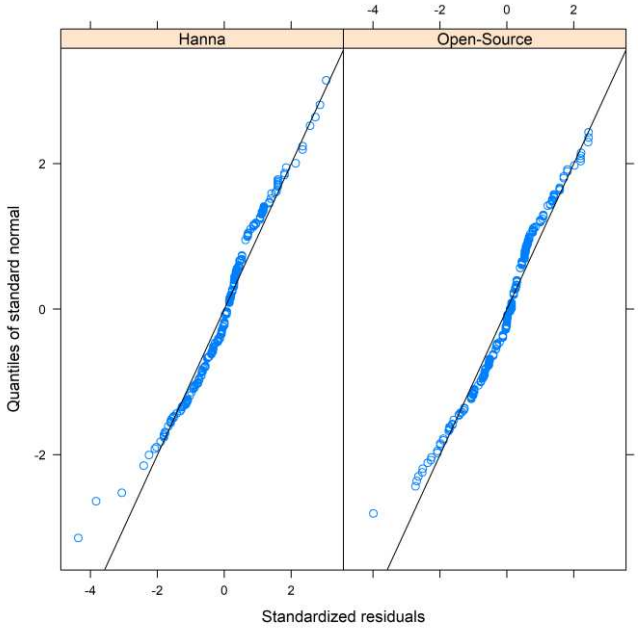


Figure A3. Standardized residuals on the horizontal axis vs the quantiles of the standard normal distribution. The plot reveals a slight deviation from the normal distribution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

Appendix B

B.1.1. Mixed-Effects models

Mixed effects model for unlinked data are described as follows [11]:

$$Y_{i1k} = b_i + b_{i1} + e_{i1k,}\, Y_{i2k} = \beta_0 + b_i + b_{i2} + e_{i2k} \quad \text{(B1)}$$

- $k = 1, \dots, m_{ij}$, are the repeated measurements.
- $i = 1, \dots, n$, is the subject's number.
- $j = 1,2$, is the method's number.
- β_0, is a fixed effect and represents the difference in the fixed biases of the methods.
- b_{ij} follow independent $N_1(0, \psi^2)$ distributions. This is an interaction term. One interpretation for b_{ij} is the effect of method j on subject i. These interactions are subject-specific biases of the methods. They are a characteristic of the method-subject combination that remains stable during the measurement period.
- b_i follow independent $N_1(\mu_b, \sigma_b^2)$ distributions.
- e_{ijk} follow independent $N_1(0, \sigma_{e_j}^2)$ distributions.
- b_i, b_{ij} and e_{ijk} are mutually independent.

To examine the measures of similarity and agreement we must retrieve the parameters of the assumed model (B1) which produces a bivariate distribution for (Y_1, Y_2). By dropping the subscripts for the sake of simplicity, we have:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_b \\ \beta_0 + \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{e_1}^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \psi^2 + \sigma_{e_2}^2 \end{pmatrix}\right)$$

Thus, $D = Y_2 - Y_1 \sim N_1\left(\beta_0, 2\psi^2 + \sigma_{e_1}^2 + \sigma_{e_2}^2\right)$

Then the model has a total of 6 unknown parameters $\left(\beta_0, \mu_b, \sigma_b^2, \psi^2, \sigma_{e_1}^2, \sigma_{e_2}^2\right)$.

Linked data are modeled as in model (B1) except for the addition of the term b_{ik}^* which represents the random effect of the common time k on the measurements.

$$Y_{i1k} = b_i + b_{i1} + b_{ik}^* + e_{i1k,}\, Y_{i2k} = \beta_0 + b_i + b_{i2} + b_{ik}^* + e_{i2k} \quad \text{(B2)}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_b \\ \beta_0 + \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_1}^2 & \sigma_b^2 + \sigma_{b^*}^2 \\ \sigma_b^2 + \sigma_{b^*}^2 & \sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_2}^2 \end{pmatrix}\right)$$

Thus, $D = Y_2 - Y_1 \sim N_1\left(\beta_0, 2\psi^2 + \sigma_{e_1}^2 + \sigma_{e_2}^2\right)$

Then the model has a total of 7 unknown parameters $\left(\beta_0, \mu_b, \sigma_b^2, \psi^2, \sigma_{b^*}^2, \sigma_{e_1}^2, \sigma_{e_2}^2\right)$.

The distribution of (Y_1, Y_2) for unlinked data is the following:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} | \tilde{u} \sim N_2\left(\begin{pmatrix} \mu_b \\ \beta_0 + \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{e_1}^2 g_1^2(\tilde{u}, \delta_1) & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \psi^2 + \sigma_{e_2}^2 g_2^2(\tilde{u}, \delta_2) \end{pmatrix}\right) \quad \text{(B3)}$$

For the linked data:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} | \tilde{u} \sim N_2\left(\begin{pmatrix} \mu_b \\ \beta_0 + \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_1}^2 g_1^2(\tilde{u}, \delta_1) & \sigma_b^2 + \sigma_{b^*}^2 \\ \sigma_b^2 + \sigma_{b^*}^2 & \sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_2}^2 g_2^2(\tilde{u}, \delta_2) \end{pmatrix}\right)$$

Based on the model parameters the heteroscedastic difference distribution is the following both for the unliked and linked data:

$$D | \tilde{u} \sim N_1\left(\beta_0, 2\psi^2 + \sigma_{e_1}^2 g_1^2(\tilde{u}, \delta_1) + \sigma_{e_2}^2 g_2^2(\tilde{u}, \delta_2)\right)$$

When the errors of models (B1) and (B2) are heteroscedastic, $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$ are replaced with $\sigma_{e_1}^2 g_1^2(u, \delta_1)$ and $\sigma_{e_2}^2 g_2^2(u, \delta_2)$. For a given u_i denoted as $\tilde{u}_i$ ($\tilde{u}$ for subject i) function g is the variance function and δ is a vector of heteroscedasticity parameters, which for $\delta = 0$ corresponds

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

to homoscedasticity. Variance covariate u is defined in advance ($\tilde{u}$) and accounts for heteroscedasticity. Choudhary and Nagaraja [11] set $\tilde{u}_i = h(\bar{y}_{i1.}, \bar{y}_{i2.}) = \bar{y}_{i1.}$ if method 1 is the reference and $\tilde{u}_i = h(\bar{y}_{i1.}, \bar{y}_{i2.}) = \frac{\bar{y}_{i1.}+\bar{y}_{i2.}}{2}$ otherwise. For the variance function g, two simple models are introduced. First, the power model, where $g(u, \delta) = |u|^{\delta}$. Second, the exponential model, $g(u, \delta) = exp(u\delta)$. The parameters δ_j can be estimated while fitting the model using ML and the "nlme" package [46]. More details on the choice of the variance function g can be found in [47]. AIC and BIC can be compared to distinguish between different candidate models for the variance functions.

Models with covariates.

Other factors might affect the agreement between the two methods. Covariates might affect the means of the methods (mean covariates), explaining part of the variability in the measurements. Covariates might also interact with the method or affect the error variance (variance covariates). In any case, the extend of the agreement is affected by the covariates. The mixed-effects models (B1) and (B2) can be extended as follows:

For unlinked data:

$$Y_{ijk} = \mu_j(x_{1i}, \dots, x_{ri}) + v_i + b_{ij} + e_{ijk} \quad \text{(B4)}$$

For linked data:

$$Y_{ijk} = \mu_j(x_{1i}, \dots, x_{ri}) + v_i + b^*_{ik} + b_{ij} + e_{ijk}$$

- $x_1, \dots, x_r$ are the mean covariates.
- v_i~independent $N_1(0, \sigma_b^2)$ and is defined as $v_i = b_i - \mu_b$
- e_{ijk}~indepentend $N_1\left(0, \sigma_{e_j}^2 {g_j}^2(u, \delta_j)\right)$,accounts for possible heteroscedasticity.
- b_{ij}~independent $N_1(0, \psi^2)$
- b^*_{ik}~independent $N_1(0, \sigma_{b^*}^2)$

Choudhary and Nagaraha [11] describe the detailed methodology for defining mean and variance -specific covariates.

B.1.2. Measurement-Error models

Taffé [29,30] and Taffé et al. [31] proposed various graphical tools to assess bias and precision of measurement methods in method comparison studies by following the measurement error model proposed by [48] but an alternative estimation procedure based on an empirical Bayes approach. The model is the following:

$$Y_{1ij} = a_1 + \beta_1 x_{ij} + \varepsilon_{1ij}, \quad \varepsilon_{1ij}|x_{ij} \sim N\left(0, \sigma_{\varepsilon_1}^2(x_{ij}, \theta_1)\right) \quad \text{(B5)}$$

$$Y_{2ij} = +x_{ij} + \varepsilon_{2ij}, \quad \varepsilon_{2ij}|x_{ij} \sim N\left(0, \sigma_{\varepsilon_2}^2(x_{ij}, \theta_2)\right)$$

$$x_{ij} \sim f_x(\mu_x, \sigma_x^2)$$

- Y_{kij} is the j^{th} replicate measurement by method k on individual i, $j = 1, \dots, n_i$, $i = 1, \dots, N$, $k = 1,2$
- n_i denotes the number or repeated measurements per subject.
- x_{ij} is a latent trait variable with density f_x representing the true latent trait.
- ε_{kij} represents measurement errors by method k.
- The variances of these methods $\sigma_{\varepsilon_k}^2(x_{ij}, \theta_k)$ are heteroscedastic and increase with the level of the true latent trait x_{ij} which depends on the vectors of unknown parameters θ_k [29].
- The mean value of the latent trait is μ_x and the variance is σ_x^2.
- It is assumed that the latent variable represents the true unknown but constant value of the trait for individual i and therefore $x_{ij} \equiv x_i$.
- The parameters a_1 and β_1 are considered fixed and the error produced by them is called systematic. Their values depend on the measurement method. Parameter a_1 is defined as the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

fixed bias (or differential bias) of the method, while β_1 is defined as the proportional bias. Fixed bias is the added constant that the measurement method adds to the true value for every measurement. Proportional bias is based on the measured quantity and is the slope of (B5). The true value is multiplied by the proportional bias and is interpreted as the amount of change in the measurement method if the true value changes by 1 unit.

This modification of the classical measurement error model considers that heteroscedasticity depends on the latent trait and not on the observed average, compared to Choudhary's and Nagaraja's methodology [11]). The model is assumed to be linear even though non-linear functions of x_i can be used and easily interpreted. To visually assess the plausibility of the straight-line model a graphical representation of $\widehat{|\varepsilon^*_{2ij}|}$ versus $\hat{x}_i$ provides a good start. Term $\widehat{|\varepsilon^*_{2ij}|}$ is the regression of the absolute values of the residuals $\widehat{\varepsilon^*_{2ij}}$ from the linear regression model $y_{2ij} = \alpha^*_2 + \beta^*_2\hat{x} + \varepsilon^*_{2ij}$ on $\hat{x}_i$ (the estimate of the latent trait) by ordinary least squares.

B.2. Indices to assess agreement and similarity.

B.2.1. Assessing Agreement.

To evaluate the agreement and similarity between the reference and the open-source device, CCC and TDI were calculated using both (B1) and (B2) [11] and methodologies provided by [34,36].

The limits of agreement [26,44,49] for both unlinked and linked data were calculated using the parameters in models (B1) and (B2) via the following formulas provided by [11]:

$$LOA = \beta_0 \pm 1.96 \cdot \sqrt{2\hat{\psi}^2 + \hat{\sigma}^2_{e_1} + \hat{\sigma}^2_{e_2}}, \text{ j=1,2} \quad \text{(B6)}$$

For heteroscedastic data $\sigma^2_{e_1}$ and $\sigma^2_{e_2}$ are replaced with $\sigma^2_{e_1} g_1^2(u, \delta_1)$ and $\sigma^2_{e_2} g_2^2(u, \delta_2)$.

For unlinked and linked data, using models (B1, B3) and following the aproach described by [11] TDI can be calculated using the following formula:

$$TDI(p) = \sqrt{\left\{\left(2\psi^2 + \sigma^2_{e_1} + \sigma^2_{e_2}\right)\chi^2_{1,p}\left(\frac{\beta_0^2}{\left\{2\psi^2 + \sigma^2_{e_1} + \sigma^2_{e_2}\right\}}\right)\right\}} \quad \text{(B7)}$$

For unlinked data, under model (B1), [11] proposed the following formula to calculate CCC:

$$CCC = \frac{2\sigma_b^2}{\beta_0^2 + 2\left(\sigma_b^2 + \psi^2\right) + \sigma^2_{e_1} + \sigma^2_{e_2}} \quad \text{(B8)}$$

While for linked data under model (B2):

$$CCC = \frac{2\left(\sigma_b^2 + \sigma^2_{b^*}\right)}{\beta_0^2 + 2\left(\sigma_b^2 + \psi^2 + \sigma^2_{b^*}\right) + \sigma^2_{e_1} + \sigma^2_{e_2}} \quad \text{(B9)}$$

Inference for TDI

Escaramis et al. [34] proposed tolerance intervals for TDI. The value k_p is obtained by replacing μ_D and σ_D by their REML estimate counterparts derived from mixed-effects models (available in [34]) in expression $\hat{k}_p = \hat{\mu}_D + z_{p_1}\hat{\sigma}_D$. For inference, a one-sided tolerance interval is computed that covers the p_1-percent of the population from D with a stated confidence.

Let T be the studentized variable of $\hat{\mu}_D + z_{p_1}\hat{\sigma}_D$. T follows a non-central Student-t distribution with non-centrality parameter $z_{p_1}\sqrt{N}$, $T \sim t_\nu\left(z_{p_1}\sqrt{N}\right)$, where $N = 2 \cdot n \cdot m$ is the total possible paired-measurement differences between the two method/devices and the degrees of freedom ν, are derived from the residual degrees of freedom. For the case of using individual-device interaction or discarding it, [34] described different cases for obtainingν. An upper bound for TDI estimate can be constructed by using the following $(1-\alpha)\cdot 100\%$ one-sided tolerance interval, where a is the type I error rate:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

$$\mathrm{UB}(1-\alpha)\cdot 100\%\left(\hat{k}_p\right)=\hat{\mu}_D+t_\nu\left(1-\alpha,z_{p_1}\sqrt{N}\right)\frac{\hat{\sigma}_D}{\sqrt{N}}$$

It corresponds to the exact one-sided tolerance interval for at least p_1 proportion of the population [50,51].

To perform a hypothesis test if the interest is to ensure that at least p-percent of the absolute differences between paired measurements are less than a predefined constant κ_0, Lin' s form of hypothesis can be followed [32].

$$H_0: k_p \geq k_0, H_1: k_p < k_0$$

H_0 is rejected at level α if:

$$\mathrm{UB}(1-\alpha)\cdot 100\%\left(\hat{k}_p\right)=\hat{\mu}_D+t_\nu\left(1-\alpha,z_{p_1}\sqrt{N}\right)\frac{\hat{\sigma}_D}{\sqrt{N}}<k_0$$

Choudhary and Nagaraja [11] use the large-sample theory of ML estimators to compute standard errors, confidence bounds and tolerance intervals. When the sample is not large enough bootstrap confidence intervals are produced. The estimators of the model-based counterparts are obtained from models (B1,B2), depending on the nature of the data (unlinked or linked respectively). To compute simultaneous confidence intervals and bounds, the percentiles of appropriate functions of multivariate normally distributed pivots are needed. Following [52] the R-package "multcomp" [53] can be used. Choudhary and Nagaraja [11] proposed a Bootstrap-t UCB and a modified nearly unbiased estimator (MNUT approximation) for computing the critical value, the p-value and the upper confidence bound (UCB)[11].

Inference for CCC

Asymptotic distribution of the estimated CCCs can be used for inference if the data are modeled via a large sample size [54]. Choudhary and Nagaraja [11] use an asymptotic distribution of the estimated CCCs to produce an upper confidence bound when the sample is large and bootstrap methods when the sample is small. Since the concordance correlation coefficient is related to the intraclass correlation coefficient (ICC), inference methods for ICC can be used for CCC [54].

B.2.2. Assessing Similarity

Following Choudhary and Nagaraja [11], to evaluate similarity the marginal distributions of Y_1 and Y_2 are examined via estimates and two-sided confidence intervals. Their distributions are given by equations (B1, B2), for unlinked and linked data. The fixed bias and the precision ratio are the two measures of similarity that will be evaluated using mixed-effects model. Last, fixed bias, proportional bias and precisions are evaluated under measurement-error models.

Fixed bias will be estimated via the model's counterparts. According to models (B1,B2) the fixed bias is estimated using: $\mu_1-\mu_2$ for unlinked and linked data.

The precision ratio is evaluated in two different cases.

First, for models that ignore subject x method interactions:

$$\lambda=\frac{\sigma_{e_1}^2}{\sigma_{e_2}^2}$$

Second, for models that include subject x method interactions:

$$\lambda=\frac{\sigma_{e_1}^2+\psi^2}{\sigma_{e_2}^2+\psi^2}$$

The precision ratios are assumed to be estimated when the errors are homoscedastic. For heteroscedastic data we replace $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$ with $\sigma_{e_1}^2\sigma_1^2(u_i,\delta_1)$ and $\sigma_{e_2}^2\sigma_2^2(u_i,\delta_2)$ thus fixed bias remains the same, but the precision ratio is given by:

$$\lambda=\frac{\sigma_{e_1}^2\sigma_1^2(u_i,\delta_1)}{\sigma_{e_2}^2\sigma_2^2(u_i,\delta_2)}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 December 2023 doi:10.20944/preprints202312.1372.v1

For inference, the method described by [48] is used for heteroscedastic data. Specifically, if θ is a vector of model's counterparts then the measure of similarity is a function of θ. Denoting the measure of similarity as φ, and b^* a value in the measurement range then $\varphi(b^*)$ is any measure of similarity in a specific value (the measure is assumed to be scalar). Substituting θ with its corresponding ML estimate, $\hat{\theta}$, in its expression gives its ML estimator$\hat{\varphi}(b^*)$. From delta method [55], when the sample size is large $\hat{\varphi}(b^*)\sim N_1\left(\varphi(b^*), G'(b^*)I^{-1}G(b^*)\right)$, where $G(b^*) = \left(\frac{\partial}{\partial\theta}\right)\varphi(b^*)|_{\theta=\hat{\theta}}$ can be computed numerically. Thus, approximate $100(1-\alpha)\%$ two-sided pointwise confidence interval for $\varphi(b^*)$on a grid of values of the measurement range can be computed as

$$\hat{\varphi}(b^*) \pm z_{1-\frac{\alpha}{2}}\{G'(b^*)I^{-1}G(b^*)\}^{\frac{1}{2}}$$

B.3. Assessing Repeatability

Following [11] and mixed-effect models, for unlinked data, repeated measurements are replications of the same underlying measurement. Instead of using the bivariate distributions (Y_1, Y_2) for measurements of the two methods on a randomly selected subject from a population, Y_j^* is defined as a replication of Y_j, where $j = 1,2$ denote the two methods/devices. By definition Y_j and Y_j^* have the same distribution. CCC and TDI are modified and are calculated. By dropping the subscripts for model (B1), for unlinked data:

$$Y_1^* = b + b_1 + e_1^*,\ Y_2^* = b + b_2 + e_2^* \tag{B10}$$

is induced, similar to (B1) by dropping the subscripts.

$$\text{Then for method 1: } \begin{pmatrix} Y_1 \\ Y_1^* \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_b \\ \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{e_1}^2 & \sigma_b^2 + \psi^2 \\ \sigma_b^2 + \psi^2 & \sigma_b^2 + \psi^2 + \sigma_{e_1}^2 \end{pmatrix}\right) \tag{B11}$$

$$\text{For method 2: } \begin{pmatrix} Y_2 \\ Y_2^* \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_0 + \mu_b \\ \beta_0 + \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{e_2}^2 & \sigma_b^2 + \psi^2 \\ \sigma_b^2 + \psi^2 & \sigma_b^2 + \psi^2 + \sigma_{e_2}^2 \end{pmatrix}\right)$$

where e_1^* and e_2^* are independent copies of e_1 and e_2 as defined in (B1).
Defining $D_j = Y_j - Y_j^*$ as the difference in two replications of method j. From (B10) and (B11):

$$D_j \sim N_1\left(0, 2\sigma_{e_j}^2\right), \text{j} = 1,2. \tag{B13}$$

Thus,

$$CCC_j = \frac{\sigma_b^2+\psi^2}{\sigma_b^2+\psi^2+\sigma_{e_j}^2} \tag{B14}$$

$$TDI_j = \left\{2\sigma_{e_j}^2\chi_{1,p}^2(0)\right\}^{\frac{1}{2}} \text{ for } j = 1,2. \tag{B15}$$

By dropping the subscripts for model (B2) for linked data:

$$Y_1^* = b + b_1 + b^{**} + e_1^*,\ Y_2^* = b + b_2 + b^{**} + e_2^*$$

$$\text{For method 1: } \begin{pmatrix} Y_1 \\ Y_1^* \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_b \\ \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_1}^2 & \sigma_b^2 + \psi^2 \\ \sigma_b^2 + \psi^2 & \sigma_b^2 + \psi^2 + +\sigma_{b^*}^2 + \sigma_{e_1}^2 \end{pmatrix}\right) \tag{B16}$$

$$\text{For method 2: } \begin{pmatrix} Y_2 \\ Y_2^* \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_0 + \mu_b \\ \beta_0 + \mu_b \end{pmatrix}, \begin{pmatrix} \sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_2}^2 & \sigma_b^2 + \psi^2 \\ \sigma_b^2 + \psi^2 & \sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_2}^2 \end{pmatrix}\right) \tag{B17}$$

Where b^{**}, e_1^* and e_2^* are independent copies of b^*, e_1 and e_2 as defined in (B2).
Defining $\mathrm{D}_j = Y_j - \mathrm{Y}_j^*$ as the difference in two replications of method j. From (B16) and (B17):

 doi:10.20944/preprints202312.1372.v1

$$D_j \sim N_1\left(0{,}2\left(\sigma_{b^*}^2 + \sigma_{e_j}^2\right)\right), \mathrm{j} = 1{,}2.$$

Thus,

$$CCC_j = \frac{\sigma_b^2 + \psi^2}{\sigma_b^2 + \psi^2 + \sigma_{b^*}^2 + \sigma_{e_j}^2} \quad \text{(B18)}$$

$$TDI_j = \left\{2\left(\sigma_{e_j}^2 + \sigma_{b^*}^2\right)\chi_{1,p}^2(0)\right\}^{\frac{1}{2}} \text{ for } j = 1{,}2. \quad \text{(B19)}$$

B.4. Recalibration Methods

For measurement-error models, a recalibration procedure is performed by computing:

$$y_{1ij}^* = \frac{Y_{1ij} - \widehat{\alpha_1^*}}{\widehat{\beta_1^*}}$$

where $\widehat{\alpha_1^*}$ is the estimate of the proportional bias and $\widehat{\beta_1^*}$ is the estimate of the fixed bias and Y_{1ij}^* is the recalibrated value. The method performs well, according to simulations, with a sample size of 100 subjects and 10 to 15 repeated measurements per individual from the reference method and only 1 from the new. It is possible that after the recalibration procedure the novel method turns out to be more precise than the reference. The recalibration procedure can be implemented using *compare_plot()* function from "*methodCompare*" package [27].

References

1. Heradio, R.; Chacon, J.; Vargas, H.; Galan, D.; Saenz, J.; De La Torre, L.; Dormido, S. Open-Source Hardware in Education: A Systematic Mapping Study. *IEEE Access* **2018**, *6*, 72094–72103, doi:10.1109/ACCESS.2018.2881929.
2. Coleman, G.; Salter, W.; Walsh, M. OpenWeedLocator (OWL): An Open-Source, Low-Cost Device for Fallow Weed Detection. *Sci Rep* **2022**, *12*, 170, doi:10.1038/s41598-021-03858-9.
3. Mesas-Carrascosa, F.J.; Verdú Santano, D.; Meroño, J.E.; Sánchez de la Orden, M.; García-Ferrer, A. Open Source Hardware to Monitor Environmental Parameters in Precision Agriculture. *Biosyst Eng* **2015**, *137*, 73–83, doi:10.1016/j.biosystemseng.2015.07.005.
4. Niezen, G.; Eslambolchilar, P.; Thimbleby, H. Open-Source Hardware for Medical Devices. *BMJ Innov* **2016**, *2*, 78–83, doi:10.1136/bmjinnov-2015-000080.
5. Jolles, J.W. Broad-scale Applications of the Raspberry Pi: A Review and Guide for Biologists. *Methods Ecol Evol* **2021**, *12*, 1562–1579, doi:10.1111/2041-210X.13652.
6. Papavlasopoulou, S.; Giannakos, M.N.; Jaccheri, L. Empirical Studies on the Maker Movement, a Promising Approach to Learning: A Literature Review. *Entertain Comput* **2017**, *18*, 57–78, doi:10.1016/j.entcom.2016.09.002.
7. Atkinson, R.K.; Sabo, K.; Conley, Q. The Participatory Web. In *Handbook of Technology in Psychology, Psychiatry and Neurology: Theory, Research, and Practice*; Nova Science Publishers, Inc., 2012; pp. 91–120 ISBN 9781621000044.
8. Louis, L. Working Principle of Arduino and Using It as a Tool for Study and Research. *International Journal of Control, Automation, Communication and Systems* **2016**, *1*, 21–29, doi:10.5121/ijcacs.2016.1203.
9. Arduino - Introduction. Available online: https://www.google.com/search?q=arduino+introduction&oq=arduino+introduction&gs_lcrp=EgZjaHJvbWUqBggAEEUYOzIGCAAQRRg7MgYIARBFGDwyBggCEEUYPDIGCAMQRRhB0gEKMTcxOTQ2ajBqNKgCALACAA&sourceid=chrome&ie=UTF-8 (accessed on 10 May 2023).
10. Oellermann, M.; Jolles, J.W.; Ortiz, D.; Seabra, R.; Wenzel, T.; Wilson, H.; Tanner, R.L. Open Hardware in Science: The Benefits of Open Electronics. *Integr Comp Biol* **2022**, *62*, 1061–1075, doi:10.1093/icb/icac043.
11. Choudhary, P.K.; Nagaraja, H.N. *Measuring Agreement*; Wiley, 2017; ISBN 9781118078587.
12. Etienne, A.; Génard, M.; Lobit, P.; Mbeguié-A-Mbéguié, D.; Bugaud, C. What Controls Fleshy Fruit Acidity? A Review of Malate and Citrate Accumulation in Fruit Cells. *J Exp Bot* **2013**, *64*, 1451–1469, doi:10.1093/jxb/ert035.

13. Tyl Catrin and Sadler, G.D. PH and Titratable Acidity. In *Food Analysis*; Nielsen, S.S., Ed.; Springer International Publishing: Cham, 2017; pp. 389–406 ISBN 978-3-319-45776-5.
14. Lado, J.; Rodrigo, M.; Zacarías, L. Maturity Indicators and Citrus Fruit Quality. *Stewart Postharvest Review* **2014**, *10*.
15. Benjamin, B.; Billah, M.; Afreh-Nuamah, K.; Obeng-Ofori, D.; Nyarko, G. Review of the Pest Status, Economic Impact and Management of Fruit-Infesting Flies (Diptera: Tephritidae) in Africa. *Afr J Agric Res* **2015**, *10*, 1488–1498, doi:10.5897/AJAR2014.9278.
16. Zheng, H.; Zhang, Q.; Quan, J.; Zheng, Q.; Xi, W. Determination of Sugars, Organic Acids, Aroma Components, and Carotenoids in Grapefruit Pulps. *Food Chem* **2016**, *205*, 112–121, doi:10.1016/j.foodchem.2016.03.007.
17. DFRobot SEN0161-V2. Available online: https://wiki.dfrobot.com/Gravity__Analog_pH_Sensor_Meter_Kit_V2_SKU_SEN0161-V2 (accessed on 10 May 2023).
18. Maxim Integrated Products DS18B20. Available online: https://www.analog.com/media/en/technical-documentation/data-sheets/DS18B20.pdf (accessed on 10 February 2022).
19. Adafruit Adalogger Featherwing Overview. Available online: https://learn.adafruit.com/adafruit-adalogger-featherwing/overview (accessed on 15 August 2023().
20. Philips Sparkfun: Nokia 5110 Lcd. Available online: https://www.sparkfun.com/products/10168 (accessed on 5 February 2021).
21. ADS1115 Ultra-Small, Low-Power, 16-Bit, Analog-to-Digital Converter with Internal Reference. Available online: https://learn.adafruit.com/adafruit-4-channel-adc-breakouts (accessed on 18 August 2023).
22. DFRobot_PH Library. Available online: https://github.com/DFRobot/DFRobot_PH (accessed 17 March 2023).
23. Bataka, E. Statistical Methods and Experimental Design for Agreement and Similarity Studies between Open-Source Software and Hardware Devices and Their Corresponding Commercials for Applications in Agriculture. Available online: https://github.com/kersee112358/Chapter-5----Ph-logger-sensor.
24. Hanna Instruments Transcat. Available online: https://www.transcat.com/media/pdf/HI9024.pdf (accessed on 20 May 2023).
25. Carstensen, B.; Simpson, J.; Gurrin, L.C. Statistical Models for Assessing Agreement in Method Comparison Studies with Replicate Measurements. *International Journal of Biostatistics* **2008**, *4*, doi:10.2202/1557-4679.1107.
26. Altman, D.G.; Blandt, J.M. Measurement in Medicine : The Analysis of Method Comparison Studiest. **1983**, *32*, 307–317, doi:10.2307/2987937.
27. Taffé, P., Peng, M., Stagg, V., & Williamson, T. MethodCompare: An R package to assess bias and precision in method comparison studies. Statistical Methods in Medical Research, 28(8), 2557–2565. https://doi.org/10.1177/0962280218759693
28. Choudhary, P. K., & Nagaraja., H. N. Measuring Agreement: Models, Methods, and Applications. Available online: https://personal.utdallas.edu/~pankaj/agreement_book/ (accessed on 14 May 2023)
29. Taffé, P. Effective Plots to Assess Bias and Precision in Method Comparison Studies. *Stat Methods Med Res* **2018**, *27*, 1650–1660, doi:10.1177/0962280216666667.
30. Taffé, P. Assessing Bias, Precision, and Agreement in Method Comparison Studies. *Stat Methods Med Res* **2020**, *29*, 778–796, doi:10.1177/0962280219844535.
31. Taffé, P.; Halfon, P.; Halfon, M. A New Statistical Methodology Overcame the Defects of the Bland–Altman Method. *J Clin Epidemiol* **2020**, *124*, 1–7, doi:10.1016/j.jclinepi.2020.03.018.
32. Lin, L.; Hedayat, A.S.; Wu, W. *Statistical Tools for Measuring Agreement*; Springer New York: New York, NY, 2012; ISBN 978-1-4614-0561-0.
33. Barnhart, H.X.; Haber, M.J.; Lin, L.I. An Overview on Assessing Agreement with Continuous Measurements. *J Biopharm Stat* **2007**, *17*, 529–569, doi:10.1080/10543400701376480.
34. Escaramís, G.; Ascaso, C.; Carrasco, J.L. The Total Deviation Index Estimated by Tolerance Intervals to Evaluate the Concordance of Measurement Devices. *BMC Med Res Methodol* **2010**, *10*, doi:10.1186/1471-2288-10-31.
35. Escaramís, G.; Ascaso, C.; Carrasco, J.L. The Total Deviation Index Estimated by Tolerance Intervals to Evaluate the Concordance of Measurement Devices, R-Script. Available online: https://static-

content.springer.com/esm/art%3A10.1186%2F1471-2288-10-31/MediaObjects/12874_2009_434_MOESM1_ESM.PDF (accessed on 10 May 2023).

36. Carrasco, J.L.; Philips, B.R., Puig-Martinez, J., King, T.S., Chinchili, V.M. Estimation of the concordance correlation coefficient for repeated measures using SAS and R. *Computer Methods and Programs in Biomedicine* **2013,** *3,* doi: https://doi.org/10.1016/j.cmpb.2012.09.002.
37. Fleiss, J. L. *The Design and Analysis of Clinical Experiments*, *1st ed.;* John Wiley & Sons, Inc, 1986.
38. Carrasco, J.L.; Martinez J.P. _cccrm: Cooncordance Correlation Coefficient. R Package Version 2.1.0. Available online: https://cran.r-project.org/web/packages/cccrm/index.html (accessed on 10 December 2023).
39. Datta, D. Blandr: Bland-Altman Method Comparison. 0.5.1. Available online: https://cran.r-project.org/web/packages/blandr/vignettes/introduction.html (accessed on 10 December 2023).
40. Grubbs, F.E. On Estimating Precision of Measuring Instruments and Product Variability. J Am Stat Assoc 1948, 43, 243, doi:10.2307/2280371.
41. Mandel, J.; Stiehler, R.D. Sensitivity--a Criterion for the Comparison of Methods of Test. *J Res Natl Bur Stand (1934)* **1954**, *53*, 155, doi:10.6028/jres.053.018.
42. Bataka, E. Statistical Methods and Experimental Design for Agreement and Similarity Studies between Open-Source Software and Hardware Devices and Their Corresponding Commercials for Applications in Agriculture. PhD thesis, University of Thessaly, Greece, 27 September 2023.
43. Khan, S.; Saultry, B.; Adams, S.; Kouzani, A.Z.; Decker, K. Since January 2020 Elsevier Has Created a COVID-19 Resource Centre with Free Information in English and Mandarin on the Novel Coronavirus COVID- 19 . The COVID-19 Resource Centre Is Hosted on Elsevier Connect , the Company ' s Public News and Information . **2020**.
44. Bland, J.M.; Altman, D.G. Measuring Agreement in Method Comparison Studies. *Stat Methods Med Res* **1999**, *8*, 135–160, doi:10.1191/096228099673819272.
45. Bongers, C.C.W.G.; Daanen, H.A.M.; Bogerd, C.P.; Hopman, M.T.E.; Eijsvogels, T.M.H. Validity, Reliability, and Inertia of Four Different Temperature Capsule Systems. *Med Sci Sports Exerc* **2018**, *50*, 169–175, doi:10.1249/MSS.0000000000001403.
46. Pinheiro J.; Bates, D.; R Core Team R Core Team (2022). _nlme: Linear and Nonlinear Mixed Effects Models_. R Package Version 3.1-162. Available online: https://cran.r-project.org/web/packages/nlme/nlme.pdf (accessed on 10 December 2023).
47. José C. Pinheiro, D.M.B. *Mixed-Effects Models in S and S-PLUS*; Springer-Verlag: New York, 2000; ISBN 0-387-98957-9.
48. Nawarathna, L.S.; Choudhary, P.K. A Heteroscedastic Measurement Error Model for Method Comparison Data with Replicate Measurements. *Stat Med* **2015**, *34*, 1242–1258, doi:10.1002/sim.6424.
49. Bland, J.M.; Altman, D.G. Agreement between Methods of Measurement with Multiple Observations per Individual. *J Biopharm Stat* **2007**, *17*, 571–582, doi:10.1080/10543400701329422.
50. Hahn, G.J. Statistical Intervals for a Normal Population Part I. Tables, Examples and Applications. *Journal of Quality Technology* **1970**, *2*, 115–125, doi:10.1080/00224065.1970.11980426.
51. Hahn, G.J.; Meeker, W.Q. *Statistical Intervals*; Wiley, 1991; ISBN 9780471887690.
52. Hothorn, T.; Bretz, F.; Westfall, P. Simultaneous Inference in General Parametric Models. *Biometrical Journal* **2008**, *50*, 346–363, doi:10.1002/bimj.200810425.
53. Multcomp. R package. Available online: https://cran.r-project.org/web/packages/multcomp/index.html
54. Carrasco, J. L., & Jover, L. (2003). Estimating the Generalized Concordance Correlation Coefficient through Variance Components. Biometrics, 59(4), 849–858. https://doi.org/10.1111/j.0006-341X.2003.00099.x
55. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer-Verlag: New York, 1998; ISBN 0-387-98502-6.