2016

# Improved interpretation of studies comparing methods of dietary assessment: combining equivalence testing with the limits of agreement

Marijka Batterham
*University of Wollongong*, marijka@uow.edu.au

Christiana van Loo
*University of Wollongong*, cmtvl646@uowmail.edu.au

Karen E. Charlton
*University of Wollongong*, karenc@uow.edu.au

Dylan P. Cliff
*University of Wollongong*, dylanc@uow.edu.au

Anthony D. Okely
*University of Wollongong*, tokely@uow.edu.au

## Publication Details

Batterham, M. J., van Loo, C., Charlton, K. E., Cliff, D. P. & Okely, A. D. (2016). Improved interpretation of studies comparing methods of dietary assessment: combining equivalence testing with the limits of agreement. The British Journal of Nutrition: an international journal of nutritional science, 115 (7), 1273-1280.

# Improved interpretation of studies comparing methods of dietary assessment: combining equivalence testing with the limits of agreement

**Abstract**

The aim of this study was to demonstrate the use of testing for equivalence in combination with the Bland and Altman method when assessing agreement between two dietary methods. A sample data set, with eighty subjects simulated from previously published studies, was used to compare a FFQ with three 24 h recalls (24HR) for assessing dietary I intake. The mean I intake using the FFQ was 126·51 (sd 54·06) µg and using the three 24HR was 124·23 (sd 48·62) µg. The bias was −2·28 (sd 43·93) µg with a 90 % CI 10·46, 5·89 µg. The limits of agreement (LOA) were −88·38, 83·82 µg. Four equivalence regions were compared. Using the conventional 10 % equivalence range, the methods are shown to be equivalent both by using the CI (−12·4, 12·4 µg) and the two one-sided tests approach (lower t=−2·99 (79 df), P=0·002; upper t=2·06 (79 df), P=0·021). However, we make a case that clinical decision making should be used to set the equivalence limits, and for nutrients where there are potential issues with deficiency or toxicity stricter criteria may be needed. If the equivalence region is lowered to ±5 µg, or ±10 µg, these methods are no longer equivalent, and if a wider limit of ±15 µg is accepted they are again equivalent. Using equivalence testing, acceptable agreement must be assessed a priori and justified; this makes the process of defining agreement more transparent and results easier to interpret than relying on the LOA alone.

**Disciplines**

Medicine and Health Sciences | Social and Behavioral Sciences

**Publication Details**

Batterham, M. J., van Loo, C., Charlton, K. E., Cliff, D. P. & Okely, A. D. (2016). Improved interpretation of studies comparing methods of dietary assessment: combining equivalence testing with the limits of agreement. The British Journal of Nutrition: an international journal of nutritional science, 115 (7), 1273-1280.

British Journal of Nutrition. Br J Nutr. 2016 Apr;115(7):1273-80

DOI: http://dx.doi.org/10.1017/S0007114516000040 , 8 pages. Published online: 16 February 2016

# Improved interpretation of studies comparing methods of dietary assessment: Combining equivalence testing with the limits of agreement .

**Short title:** Equivalence for agreement of dietary methods

**Keywords:** equivalence, agreement, dietary assessment, Bland and Altman

**Authors:**

Marijka J Batterham[1]

Christel Van Loo[2]

Karen E Charlton[3]

Dylan P Cliff[2]

Anthony D Okely[2]

1. National Institute for Applied Statistics Research Australia, University of Wollongong, Australia
2. Faculty of Social Sciences, University of Wollongong, Australia
3. Faculty of Science, Medicine and Health, University of Wollongong, Australia

**Corresponding Author:**

Marijka Batterham, National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Northfields Ave Wollongong NSW 2522, Australia

Ph 61 2 4221 8190 email: marijka@uow.edu.au

## Abstract

The aim of this paper is to demonstrate the use of testing for equivalence in combination with the Bland and Altman method when assessing agreement between two dietary methods. A sample dataset, with 80 subjects simulated from previously published research, is used to compare a food frequency questionnaire with three 24 hour recalls for assessing dietary iodine intake. The mean iodine intake using the FFQ was 126.51µgSD54.06 and using the three 24 hour recalls 124.23µgSD48.62. The bias and 90% CI were -2.28µgSD43.93 with a 90%CI-10.46, 5.89µg and a limits of agreement -88.38, 83.82µg. Four equivalence regions were compared. Using the conventional 10% equivalence range the methods are shown to be equivalent both by using the CI (-12.4, 12.4µg) and the two one sided tests approach (lower t=-2.99(79df)P=0.002, upper t=2.06(79df)P=0.021). However we make a case that clinical decision making should be used to set the equivalence limits and for nutrients where there are potential issues with deficiency or toxicity a stricter criteria may be needed. If the equivalence region is lowered to ±5µg, or±10µg, these methods are no longer equivalent, and if a wider limit of ±15µg is accepted they are again equivalent. Using equivalence testing acceptable agreement must be assessed a priori and justified, this makes the process of defining agreement more transparent and results easier to interpret than relying on the limits of agreement alone.

## Introduction

The Bland and Altman (BA) method[(1)] has been routinely used for assessing relative agreement between two dietary methods. The rationale for doing this, typically, is that while the reference method, or gold standard, is deemed to be more accurate, it also has substantial participant burden to complete and resources to analyse. Often, the Food Frequency questionnaire (FFQ) method is compared against Food Records, either weighed or unweighed, or repeated 24hr dietary recalls. Food frequency questionnaires are easier to implement, less burdensome for participants to complete and less costly to analyse[(2)]. It is necessary to demonstrate that FFQ results are equivalent to a reference method before it can be used with confidence. Interpretation of the results of the BA method is straightforward when it is clear that the methods do not agree. In practice, this is defined by a large and statistically significant bias using a dependent samples test (paired t test or Wilcoxon matched pairs test). However, difficulty arises in determining equivalence of two dietary methods when they are shown by the BA method to be in agreement. For example a bias of 837kJ with a limits of agreement of -5192kJ to 6865kJ was defined as “reasonably acceptable” agreement in a study[(3)] comparing a FFQ to an estimated food diary. Likewise, compared to a 24 hour recal a FFQ was reported to have a bias of 1091kJ with a limit of agreement of -2792kJ to 4974kJ[(4)]. This was described as “performing well”and was considered to have “fair” or “adequate” agreement despite the large and statistically significant bias and wide limits of agreement. These two examples demonstrate a lack of consideration on what constitutes a clinically acceptable difference between dietary methods. The limits of agreement in these studies encompass a range of intake between 7766kJ and 12056kJ, which is the magnitude of intake that represents the entire recommended daily intake for an adult (i.e. 8400 to 11700kJ[(5; 6)]). This is clearly undesirable, yet appears to be the current practice in the published nutrition literature. As Bland and Altman, themselves stated, “How far apart measurements can be without leading to problems will depend on the use to which the result is put, and is a question of clinical judgement. Statistical methods cannot answer such a question”[(7)].

The aim of this paper is to consider how two methods can be demonstrated as being equivalent when the BA indicates agreement. Here, we make a case for combining formal testing of equivalence with the BA method for assessing agreement between methods. Performing a test of equivalence requires an a priori assessment of what constitutes a clinically acceptable difference between two methods. In this paper, we first consider how agreement is described in the nutrition literature for validation of FFQs using the BA method. Secondly, we compare the use of equivalence testing to the BA method for assessing agreement between two methods using an original dataset. The emphasis of this paper is to demonstrate the need to be able to accurately define what constitutes clinical agreement - before being able to interpret the level of agreement between these methods - and to encourage the use of both methods in validation studies.

## Methods

To identify a sample of FFQ validation literature describing agreement using the BA method, a search of the database Web of Science (accessed 20th March 2015) was conducted. This search returned 24847 citations for the initial Bland and Altman paper, of which 250 were identified under the sub search for Food Frequency Questionnaire. We then selected the 10 papers with the highest number of citations, available through our institutional subscriptions, which aimed to validate an FFQ using the BA method.

To demonstrate equivalence testing and compare this with the BA method, a dataset consisting of a random sample of 80 participants was simulated based on a previously published analysis using the means of iodine intake assessed using the average of three repeated 24hr recalls and a FFQ (3x24hrR 118.88µg SD48.95, FFQ 120.19µg SD55.98 and correlation 0.614; P<0.001)(8). The dataset was simulated using the matrix and drawnorm commands in STATA (Version 12, STATA Inc, College Stn, Tx). Simulated data was chosen, instead of the actual data, in this example to allow data sharing without any ethical considerations. In addition the initial dataset was right skewed and transformed for analysis and the simulated data is normally distributed to assist with interpretation.

The agreement of methods was interpreted using both a BA limits of agreement and an equivalence approach. Both methods advocate acceptance on the basis of a clinical decision, however in the case of the equivalence approach this must be explicitly stated a priori(9).

The BA method(1) involves plotting the difference between the two methods against the average of the two methods and examing the mean bias, determining the 95% CI of the bias and any trend in the bias. The precision of the limits are rarely considered in interpreting the BA plot. Interpreting the precision of the limits involves calculating and interpreting the 95% CI of the upper and lower limit and is detailed with an example in the initial Bland and Altman paper(1). Further reference to this on Martin Bland's website (https://www-users.york.ac.uk/~mb55/meas/sizemeth.htm accessed 28th August 2015) demonstrates clearly the effect of sample size on these estimates and emphasises that it is important not only to consider the width of the limits of agreement (LOA), but also the precision with which these have been estimated.

Equivalence testing was performed using the two one sided test (TOST) procedure(10) and also by using the confidence interval approach(11). Both are valid approaches and the use of one or the other depends on whether there is a preference for the use of a P value or CI. Equivalence testing is widely used in the pharmaceutical industry where a new drug, which may have fewer side effects or be less costly to produce, is compared to the standard drug to determine if the therapeutic effect is equivalent within a predefined range(12). If differences in means ($d$) are considered using a paired t test, as in the traditional framework, the intention is to demonstrate that a new drug or method is different (generally with the aim of showing superiority). In this case, the null hypothesis states that there is no difference between the treatments, while the alternate hypothesis states that there is a difference. Based on this paradigm, established by Neyman and Pearson(13), it can only be demonstrated that $d \neq 0$, or that there is insufficient evidence to demonstrate that $d \neq 0$. What cannot be demonstrated is that $= 0$ , that is the null hypothesis cannot be proven. With a small sample size, it is difficult to show that $d \neq 0$ and an erroneous conclusion that there is no difference (type 2 error) may be made, particularly if the difference is small and the variance is large(14). In this situation we may conclude that the two methods agree as we do not have adequate power to demonstrate the difference is statistically significant. Alternatively for every $d$ there is a sample size where it can be demonstrated that $d \neq 0$, regardless of whether this difference has any practical meaning. In this situation we may conclude that the methods do not agree when the difference between them is actually to small to have any clinical meaning. Thus, the statistical significance is unrelated to the practical or clinical significance. When demonstrating equivalence, the hypotheses are reversed such that the null states that there is a difference ($H_0$: $|d| \geq \Delta$ , where $d$ is the difference between the methods and $\Delta$ is the prespecified equivalence interval) and the alternative hypothesis is that of no difference ( $H_a$: $|d| < \Delta$) (15). Equivalence trials require an a priori specification of an acceptable equivalence range. Determination of this range needs to be guided by clinical acceptability of the range of measures. Wellek discusses arbitrary ranges when the equivalence range is unknown(9) and other arbitrary decisions such as ±10% of the reference mean have been employed in the literature on physical activity(16). In general this equivalence region is poorly defined. A review of 332 non inferiority and equivalence pharmaceutical trials found that half of these considered 0.5SD or less of the difference between treatments to be an “irrelevant” difference(17). While TOST is not the most powerful equivalence test(9; 18), it’s relative ease of use and interpretation(15) make it the preferred approach for nutritional applicatons.

Both the BA and Equivalence approaches are most easily interpreted visually. In our analysis, we present the traditional BA plots with the equivalence intervals incorporated. The figures contain the equivalence interval, as well as the 90% CI of the difference and the limits of agreement. These figures can be plotted easily in most statistical packages or in Microsoft Excel. This approach is adapted from the one proposed in the SAS macro “Concord” which presents a BA style plot, that incorporates the equivalence interval and 90%CI instead of the LOA(19), and we also present the results as confidence interval plots and in tabular form to show different options of presentation.

Given that we wish to provide practical guidelines on the conduct of equivalence tests we consider their use in STATA (V12, StataCorp LP, College Station Tx), SAS (V9.3 SAS Inc, Cary NC), SPSS (V21, IBM Corporation Armonk NY) and R(V3.2.1 www.cran.r-project.org(20)) and instruction on the use of each of these is considered in Appendix A. In this example, we considered four regions of equivalence to demonstrate the proposed methodology and the differences between equivalence and non equivalence. The four equivalence regions chosen for this example demonstrate how to interpret clear equivalence, non equivalence and an intuitively ambiguous result.

## Results

A summary of the ten validation studies identified from the literature review is provided in Table 1. Only three of the ten papers considered a priori what an acceptable difference between the methods would be, while none of the authors discussed what was considered an acceptable LOA. All papers reported and discussed the correlation coefficient as a method of establishing validity, although two discussed the limitations of this approach. In most cases the results were compared only to other literature and no clinically defined or practical implications of the LOA were discussed. Seven of the ten studies performed hypothesis testing (Wilcoxon, paired t test) to determine if the mean difference between the methods was statistically significant.

Table 2 presents the results of the Bland and Altman comparisons and the equivalence tests for the simulated data in tabular form. Figure 1 presents the Bland and Altman plots with the equivalence intervals and 90%CI of the difference. Figure 2 presents the confidence interval plot, Figure 2a shows a confidence interval plot expressing the x axis as the difference between the two means, as is the traditional approach used for pharmaceutical trials. Figure 2b shows a confidence interval plot expressed relative to the mean intake of iodine using the 3x24hrR, the two plots (2a and 2b) are identical in interpretation, in this case the methods are equivalent if the 90% CI is contained within the prespecified equivalence region. All equivalence methods show that the FFQ is only considered to be equivalent to the 3x24hrR when the equivalence margin is set at 10% of the mean of the 3x24hrR (12.24µg), or alternatively at 15µg. The methods are not equivalent when the margin is set at 5µg. The methods are also not equivalent when the margin is set at ±10µg because although the mean difference meets the criteria on one side (the upper 90% CI being 5.89 which is within the upper bound of 10µg), the lower bound is outside the range (-10.49µg < 10µg) and both sides must be within the region to meet the assumption of equivalence. This is also reflected in the P values, both of which must be significant for equivalence to hold. Commands and outputs for the tests in SAS, R, STATA and SPSS are shown in Appendix 1. Figure 3 shows the BA LOA plot with the 90% CI of the mean bias used for the equivalence testing and the 95% CIs of the upper and lower limits of agreement (numerically represented in Table 2).

## Discussion

This paper demonstrates the use of assessing equivalence in dietary studies that compare two methods for agreement. Equivalence is presented to be used in conjunction with the more commonly applied Bland and Altman limits of agreement method. The advantage of the equivalence method is that it requires the clinican to make an *a priori* assessment of what represents agreement, rather than accepting or rejecting the LOA determined in the Bland and Altman analysis *a posteri*. The equivalence approach can be assessed using confidence intervals, either independently or in combination with a BA plot, or equivalence can be assessed in the traditional paradigm of P values using two one sided tests (TOST).

Frequently, the agreement between two dietary methods is assessed using the Bland and Altman analysis, and the decision whether or not to determine agreement is based on a dependent samples test (paired t test or Wilcoxon matched pairs test). This approach was not advocated by Bland and Altman and their initial paper that describes the method makes no reference to hypothesis testing regarding the bias. Rather, the initial manuscript by Bland and Altman(1) states "How far apart the measurements can be without causing difficulties will be a question of judgement. Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size(1)."

Discussion to date on what constitutes a clinical LOA in the nutrition literature is limited. For example in the studies reviewed here many compared their LOA to other studies(21) but still with no discussion on whether this was acceptable in practice. In addition, when assessing agreement the 95% CIs of the limits (as shown in Fig 3) are rarely considered. These can be wide particularly for small datasets and should be reported, discussed and considered particularly when estimating sample sizes, as advocated in the early BA literature. When only considering the LOA themselves we may be prepared to accept that the measures agree however the interpretation of the 95% CIs of the LOA suggests that we could have an upper LOA as high as 117.96µg or a lower LOA as low as -122.24µg with repeated sampling.

Judging what is an acceptable equivalence between two methods is not a trivial procedure.[17; 22] Even in the pharmaceutical domain where equivalence tests are most often used, a systematic review found that only 134 of 314 studies provided a rationale for the difference used[17]. Given the number of agreement studies published in the field of nutrition, it is necessary to be able to determine the clinical, rather than just the statistical interpretation, of the results.

The question of what constitutes equivalence in the field of nutrition is complex. This may differ, depending on the nutrient being assessed and the population that is being studied. In the case of iodine, the estimated average requirement reported in the Australian Nutrient Reference Values is 100µg/day for adults, with a recommended daily intake of 150µg/day and an upper limit of 1100µg/day[23]. Estimated average intakes in the Australian population based on the most recent (2011-12) Australian nationally representative Health Survey were 191µg in males and 152µg in females[24]. Therefore, for the general population, a 10% equivalence based on the mean of the reference food record appears reasonable. In populations where intakes may be inadequate (for example pregnant women[25; 26] and where the consequences of inadequacy have serious impacts on health outcomes, more stringent equivalence limits may be warranted.

Consideration of why it is important to state the acceptable LOA or equivalence a priori is warranted. While there was a large range of sample sizes in the studies presented here (n=61-785) these were selected as being the most cited and dietary validation studies can be conducted with relatively small sample sizes (for example n=49[27]). This may lead to an erroneous acceptance of the null hypothesis due to limited power to detect a difference in the traditional hypothesis test (that is a type 2 error). In our particular example, the power to detect a mean difference of 2.28µg(SD43.93) with 80 subjects is only 0.084. In order for this difference (2.28µg with an SD of 43.93µg) to be statistically significant at an alpha of 0.05 with 80% power, a sample size of 2112 would be required. As the sample size increases, the probability of rejecting the null hypothesis in the traditional null hypothesis testing framework increases, while smaller sample sizes result in the opposite trend[15].

In this example, we consider only the paired t test for equivalence as it is generally the case that two dietary assessment methods would be compared on the same subjects. Independent t test methods also exist for both normally distributed and non normal data. It is often the case that dietary intake data is skewed as was the case with the initial dataset on which the simulation used in the present an analysis was based[8]. Non normal data can be analysed using a similar approach for either paired or independent data based on the robust t test of Yuen[28] in the “equivalence” package in R. Log transformation can also be considered. In this case the interpretation relies on backtransformation and the results represent the ratio of the two methods, generally expressed as a percentage with absolute equivalence being 100%. SAS has a `“dist=lognormal”` option in PROC TTEST where the TOST procedure is conducted which will convert output and produce data based on the geometric (or backtransformed) mean. When backtransforming logarithmic data, a difference of ±10% is approximately symmetrical, however wider limits will not be. For example if the equivalence region chosen is ±20% this will correspond to a range of 80% to 125%

when the ratio is backtransformed. This relationship must be considered when setting equivalence limits with log transformed data. Log transformations are commonly used in pharmaceutical equivalence testing and these concepts have been covered in the related literature(12). The equivalence approach can also be applied to other hypothesis testing such as equivalence of slopes or trend(29). In addition, multisample and multivariate tests have also been described but are beyond the scope of what is covered here.

This paper is designed to assess methodological comparison studies based on agreement using an example based on our previous research. There are other methods for judging the usefulness of new dietary assessment tools such as the method of triads which we have employed previously(8) or missclassification but they are not discussed here. Lombard and colleagues(30) provide a recent review and recommendations on the use of other methods, specifically applicable to nutrient assessment. The comparison of 3x24hrR to an FFQ outlined here is an example of an approach which can be applied not only to dietary methodology but to other methods used in nutrition practice and research which are commonly assessed for agreement using Bland and Altman methodology. These include comparing resting energy expenditure prediction equations to indirect calorimetry(31; 32; 33; 34), bioelectrical impedance analysers to dual energy x-ray absorptiometry for assessing body composition(35; 36; 37; 38) and in validating physical activity assessment tools(39; 40; 41).

In summary we have introduced here an equivalence approach to be used in conjunction with the Bland and Altman method in order to encourage clinicians to establish up front what constitutes a clinically meaningful difference between the two methods being considered. This not only makes interpretation of the results of the study clear but also assists with assessing the necessary sample size in planning the study.

**Conflict of interest**

None

**Financial support**

None

**Authorship**

The research idea was conceived during discussions between MB, CVL, DC and AO. MB formalised the study idea, simulated and analysed the data and drafted the paper. CVL assisted in study conception and comparing the analytical approaches and reviewing the manuscript. DC assisted in the study conception and reviewed the manuscript. KC assisted with the provision of the context and practical applications of the iodine example. AO provided constructive feedback on the study design and implications and reviewed the manuscript.

Table 1. Summary of a highly cited sample of the literature assessing agreement of an FFQ with a reference method using the Bland and Altman method.

| **Paper and comparator method** | **Sample size** | **Correlation performed to assess validity** | **Significance test of differences** | **BA plots presented** | **A priori assessment of acceptable agreement** | **Conclusion** | **Justifcation of conclusion** | **Number of citations*** |
|---|---|---|---|---|---|---|---|---|
| Villegas et al 2007[21] FFQ, 24HR | 195 | yes | Wilcoxon, all P<0.05. | Energy, protein, fat, carbohydrate, 8 nutrients not shown | 100% of ratio of FFQ/24 FR, no LOA | “good agreement” | Compared with other studies. | 64 |
| Matthys et al 2007[42] FFQ/EFD | 104 | Yes | Wilcoxon, | No BA plots, mean difference and SD presented | Nil | “acceptable for assessing population median intakes” for some food groups. LOA “broad for all food groups” | Acceptability based on P≥0.05 in the Wilcoxon tests | 43 |
| Weber Cullen et al 2007[43] FFQ/24HR | 83 | Yes | Paired t tests | No BA plots, mean difference and SD presented | Nil | BA plots done however not shown as they “indicated no association between the difference and the mean of the two measures”. Validity for some nutrients, but not most food groups in adolescents based on t test. | Based on paired t tests | 65 |
| Hjartaker et al 2007[4] FFQ/24HR | 238 | Yes | Wilcoxon | Energy, fibre, alcohol, 20 nutrients not shown, however the 3 figures were shown to represent 3 trends observed in the nutrients assessed. | Nil | LOA wide. Most means lower in FFQ. Three general trends in the bias over the range measured. | Compared with other studies. “overall relative validity …comparable to that of FFQs used in other large cohorts often described as ‘fair’ or ‘adquate’”. | 48 |
| Brantsæter et al 2008[44] FFQ/WFD | 119 | Yes | Wilcoxon | Energy, Fruit/vegetable/juice(g/day), 24 nutrients/vitamins/elements not shown as “the plots were similar to the plot of energy intake”. | Nil | “Bias was small, whereas the confidence limits were wide”<br>“..produces realistic and relatively precise estimate of habitual intake of energy” | Based on small bias. | 85 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Toft et al 2008[45] FFQ/DH | 264 | Yes, although limitations discussed. | Wilcoxon by sex, 38/42 P<0.05 | Saturated fat, Energy, 19 nutrients not shown | Nil | "acceptable agreement" with "small differences". Saturated fat tendency for increased underreporting with increased intake. | No definition of acceptable agreement given. | 35 |
| Watson et al 2009[46] FFQ/AFR | 113 | Yes | No | B-carotene, calcium, 20 nutrients not shown | Good (difference ≈1SD mean reference)<br>Fairly good (difference ≈2SD mean reference)<br>Poor (difference ≈3SD mean reference) | "Not suitable for estimating absolute agreement" | Positive differences,<br>Wide limits of agreement, increasing difference with increasing bias. | 59 |
| Zhang et al 2009[47] FFQ/EFD | 61 | Yes | Wilcoxon, however results not reported | Energy, Vitamin C, 12 nutrients not shown | 100% of ratio of FFQ/24 FR, no LOA | Differences stated no discussion on whether they agree with 100%, LOA similar or narrower to other studies. State BA plots show no linear trend and there is "reasonable comparative validity" | Compared with other studies. | 42 |
| Ambrosini et al 2009[48] FFQ/EFD | 785 | Yes, although discussed limitations | No | Energy, carbohydrate, protein, fat, 18 nutrients not presented. | Nil | "The majority of nutrients showed average agreement that was significantly different from 100%" however the criteria for significance was not stated. 95% CIs are presented. | "Most LOA ranged from 50-250%" similar to other studies. | 39 |
| Fernández-Ballart et al 2010[3] FFQ/EFD | 158 | Yes | No | Vegetable, meat/meat product, potato, legumes, energy, protein, thiamine, cobalamin presented as examples showing no trend or systematic trend over range of values, 26 nutrients/foods not shown | Nil | "agreement between these two methods was also reasonably acceptable". "It was found to have acceptable levels of reproducibility and validity" | Compared to other studies, expected general overestimation by FFQ[49]. | 126 |

* number of citations on Web of Science on the 28th August 2015.

WFD weighed food diary, EFD estimated food diary 24HR, 24 hour recall, DH diet history, AFR assisted food record.

**Table 2**. Summary statistics, paired t test, Bland and Altman LOA and equivalence tests for assessing agreement between the 3 x 24hrR and the FFQ.

| Method (n=80) | Iodine (mean) | SD | Minimum | Maximum |
|---|---|---|---|---|
| 3x24hrR | 124.23 µg | 48.62 µg | 29.61µg | 240.00µg |
| FFQ | 126.51 µg | 54.06 µg | 13.03µg | 244.82µg |

| | | | | | | |
|---|---|---|---|---|---|---|
| Paired t test | | | | | | |
| Mean difference 3x24hrR -FFQ | SD | SEM | 95% Confidence Interval of the Difference | | t (df=79) | P |
| -2.28µg | 43.93 µg | 4.91 µg | -12.06 µg | 7.49 µg | -.465 | .643 |
| Bland and Altman Limits of Agreement | | | | | | |
| BA bias | SD | SEM | Limits of Agreement | | 95% CI of lower limit | 95% CI of upper limit |
| -2.28 µg | 43.93 µg | 4.91 µg | -88.38 µg | 83.82 µg | -122.24 µg,<br>-54.52 µg | 49.96 µg,<br>117.96 µg |
| Paired equivalence test | | | | | | |
| Mean difference 3x24hrR -FFQ | SD | SEM | 90% Confidence Interval of the Difference | | t (df=79) | P |
| -2.28 µg | 43.93 µg | 4.91 µg | -10.46 µg | 5.89 µg | | |
| Equivalence region -2.28±5 µg iodine | | | | t upper | 0.55 | 0.291 |
| -5 > -10.46, 5.89 > 5 DECISION: NOT EQUIVALENT | | | | t lower | -1.48 | 0.071 |
| Equivalence region -2.28±10 µg iodine | | | | t upper | 1.57 | 0.060 |
| -10 > -10.46, 5.89 > 10 DECISION: NOT EQUIVALENT | | | | t lower | -2.50 | 0.007 |
| Equivalence region -2.28±15 µg iodine | | | | t upper | 2.59 | 0.006 |
| -15 > -10.46, 5.89 > 15 DECISION: EQUIVALENT | | | | t lower | -3.52 | 0.000 |
| Equivalence region -2.28±10% (±12.4) µg iodine | | | | t upper | 2.06 | 0.021 |
| -12.4 > -10.46, 5.89 > 12.4 DECISION: EQUIVALENT | | | | t lower | -2.99 | 0.002 |

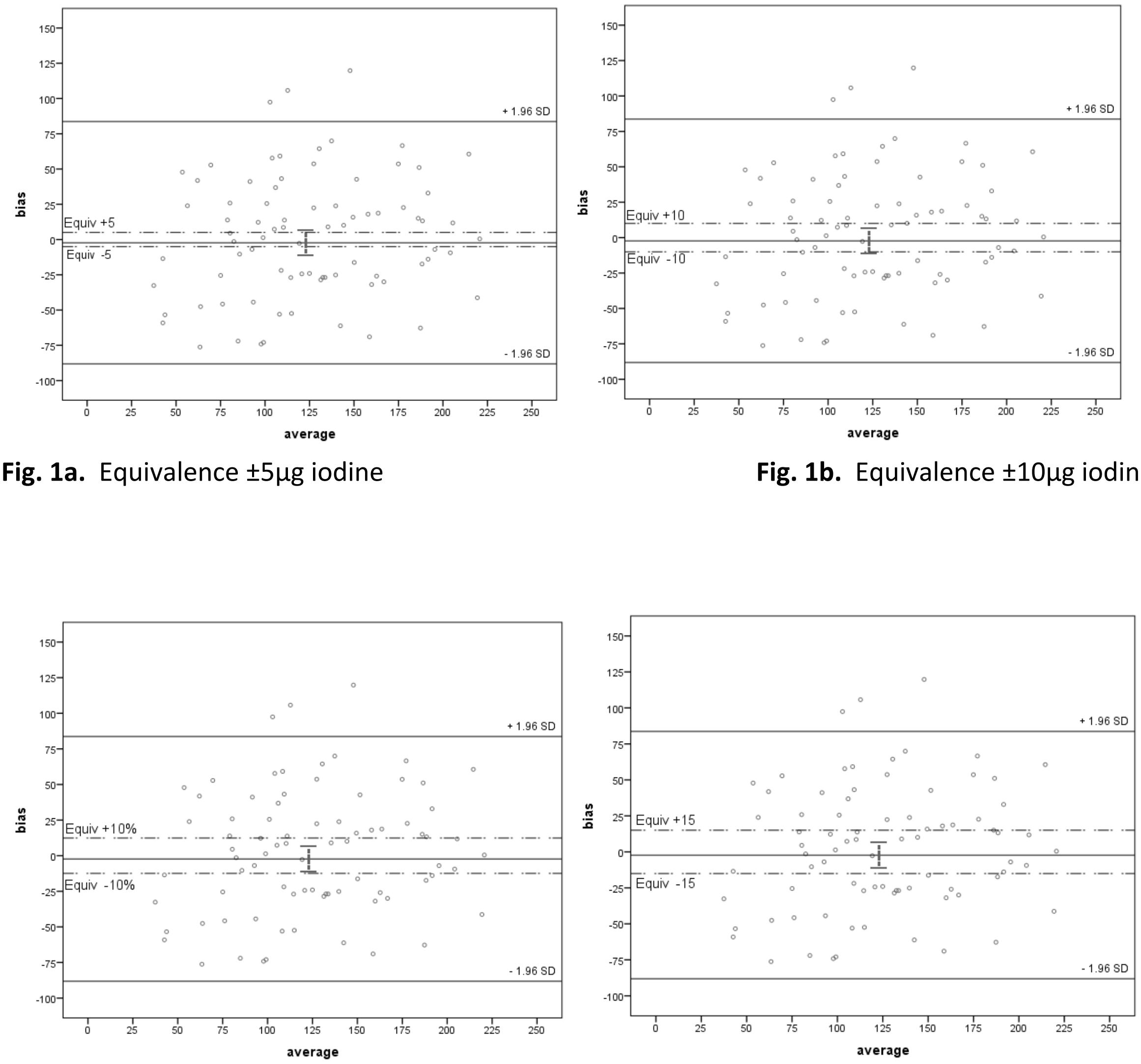


**Fig. 1a.** Equivalence ±5µg iodine

**Fig. 1b.** Equivalence ±10µg iodine

**Fig. 1c.** Equivalence ±10% mean iodine 3x24hrR

**Fig. 1d.** Equivalence ±15µg iodine

**Fig. 1.** Bland and Altman plots with superimposed equivalence intervals and the 90%CI of the mean difference. 3 x 24hrR; average of 3 24 hour recalls, FFQ; food frequency questionnaire

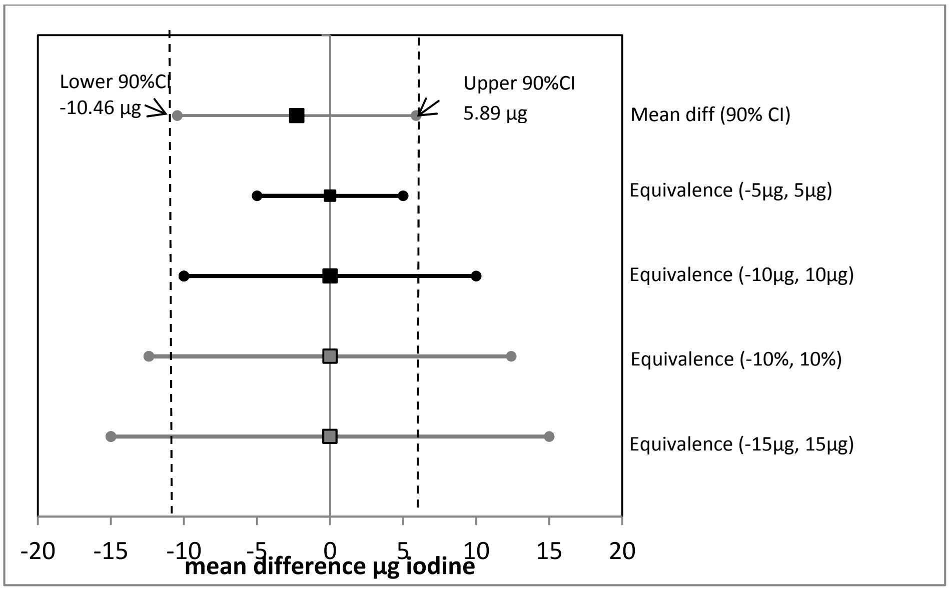


**Fig. 2a.** Confidence interval plot using the mean difference between the 3x24hrR and FFQ

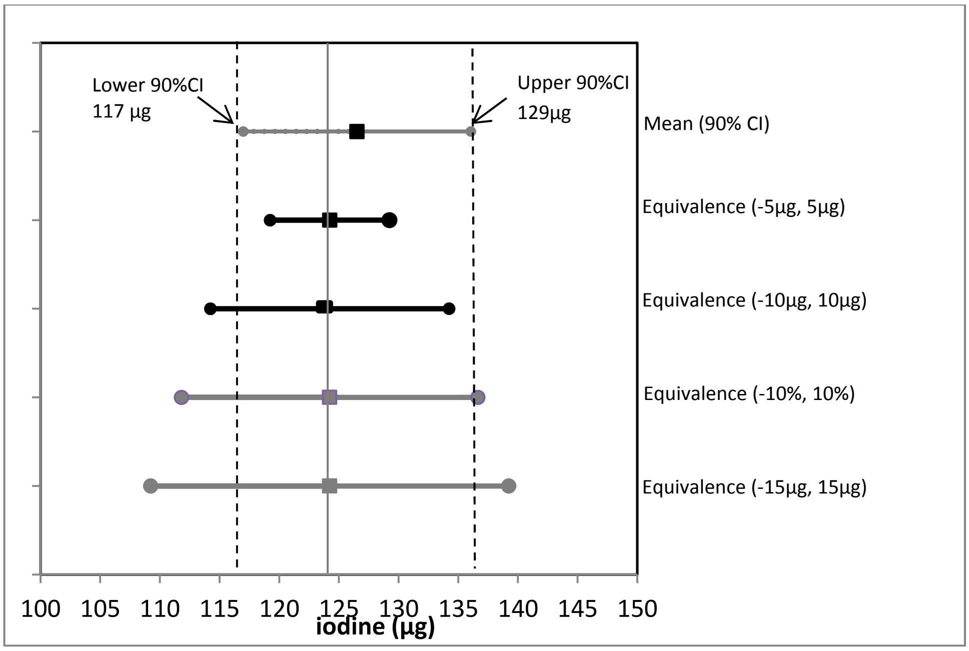


**Fig. 2b.** Confidence interval plot using the mean iodine intake in the 3x24hrR (124.23µg). 3 x 24hrR; average of 3 24 hour recalls, FFQ; food frequency questionnaire

**Fig. 2.** Confidence interval plots

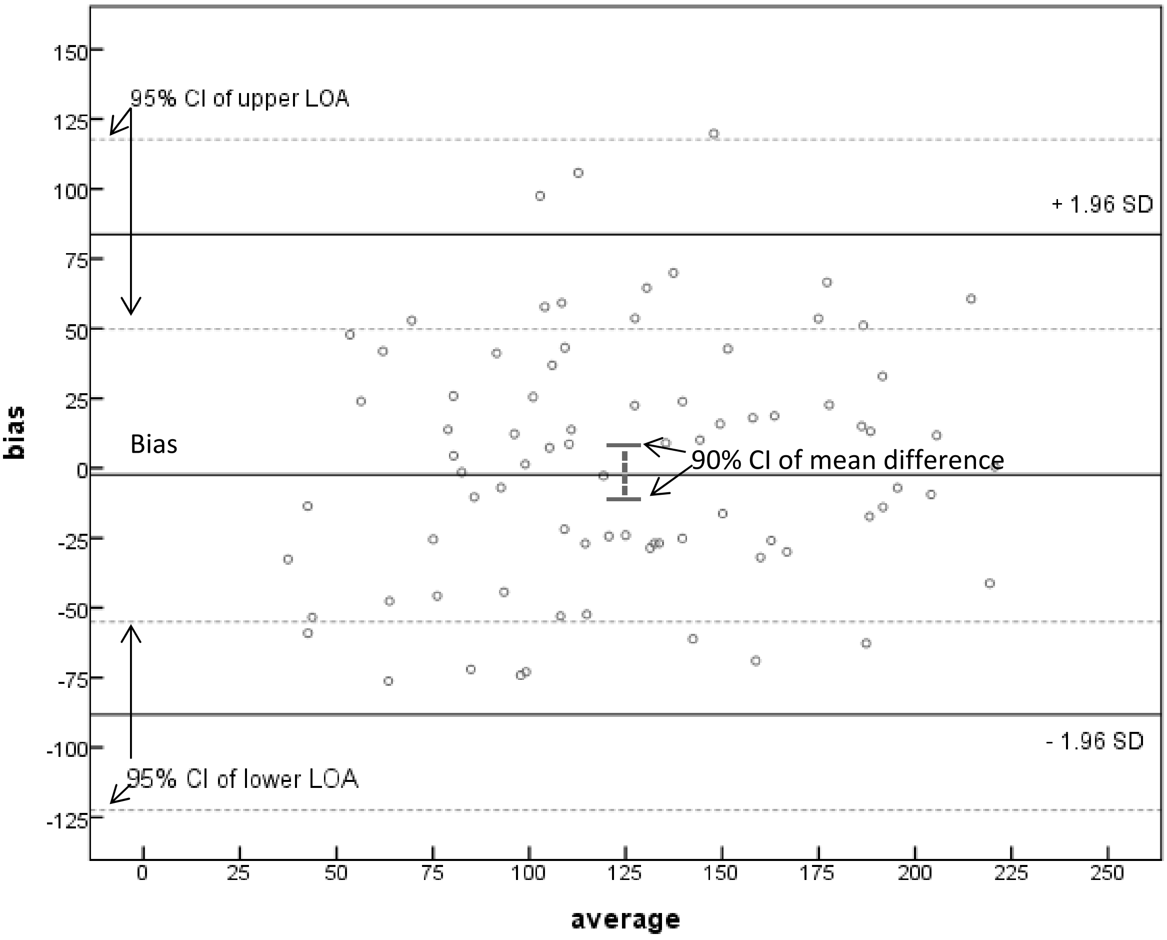


**Fig. 3.** 95% CI of the upper and lower Limits of Agreement for the mean bias in iodine intake (µg) between the 3 x 24hrR and the FFQ

**Appendix 1, Supplementary online material.**

Conducting the Equivalence test in different statistical packages. In these examples, equivalence at 5 and 15µg are shown to demonstrate the contrast output when the methods are determined to be equivalent and when they are not equivalent. The dataset is also available on request (marijka@uow.edu.au) to replicate the analyses.

1. R
   In R equivalence testing can be conducted easily using the package “equivalence” (www.cran.r-project.org/web/**packages/equivalence/equivalence**.pdf)

The tost command (in bold) can be used for paired or independent data, by specifying a single variable the paired test is used. The test is conducted on the bias, the difference between the methods. In this analysis this variable is called “bias” and the dataset is called “iodine”. The bias is the difference between the x24hrR (average of 3 24 hour recalls) and the FFQ (food frequency questionnaire)

```
tost(iodine$bias, y=NULL, alpha=0.05, epsilon=5)
$mean.diff
[1] 2.284939

$se.diff
[1] 4.91133

$alpha
[1] 0.05

$ci.diff
[1] -5.889338 10.459215
attr(,"conf.level")
[1] 0.9

$df
df
79

$epsilon
[1] 5

$result
[1] "not rejected"

$p.value
[1] 0.2909752

$check.me
[1] -0.430123  5.000000
attr(,"conf.level")
[1] 0.4180496
```

*The P value is >0.05 (0.2909752) and indicates the methods are not equivalent.*

```
tost(iodine$bias, y=NULL, alpha=0.05, epsilon=15)
$mean.diff
[1] 2.284939

$se.diff
[1] 4.91133

$alpha
[1] 0.05

$ci.diff
[1] -5.889338 10.459215
attr(,"conf.level")
[1] 0.9

$df
df
79

$epsilon
[1] 15

$result
[1] "rejected"

$p.value
[1] 0.00572852

$check.me
[1] -10.43012  15.00000
attr(,"conf.level")
[1] 0.988543
```

*The P value is <0.05 (0.00572852) and indicates the methods are equivalent.*

2. In SAS V9.3 (SAS Inc, Cary NC), equivalence testing is available through the PROC TTEST procedure, in SAS it is necessary to specify that the test is paired and the values for the FR and FFQ are used. SAS also produces graphical output. As mentioned in the text, the SAS macro concord(19) produces equivalence tests and Philip Dixon(29) provides syntax on determining equivalence using the PROC MIXED procedure in an online archive EquivSlope.sas http://www.esapubs.org/archive/ecol/E086/094/suppl-1.htm

```
Proc ttest data=iodine tost(-5,5);
paired FR*FFQ;
run;
```

The SAS System

The TTEST Procedure

Difference: FR - FFQ

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 80 | -2.2849 | 43.9283 | 4.9113 | -119.8 | 76.2301 |

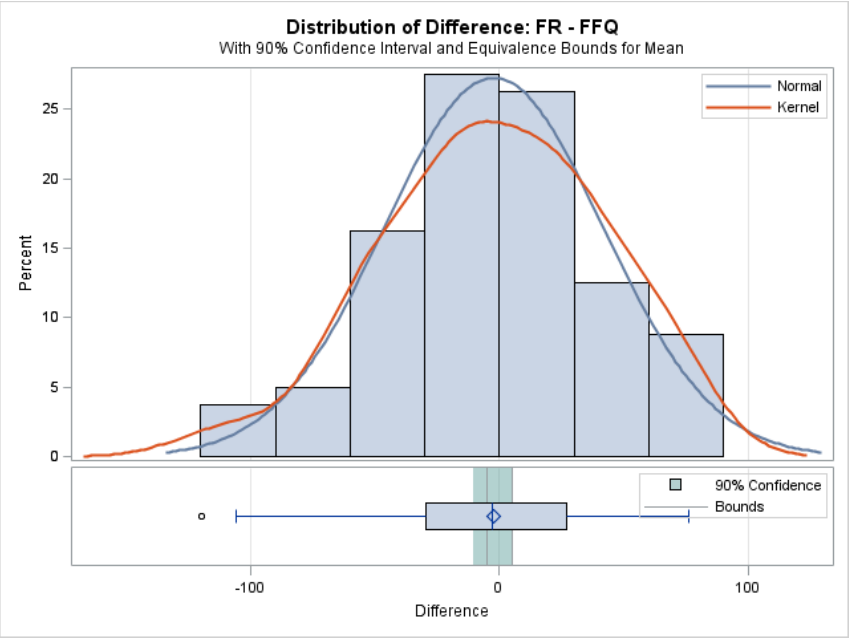
| Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|
| -2.2849 | -12.0607 | 7.4908 | 43.9283 | 38.0178 | 52.0318 |

TOST Level 0.05 Equivalence Analysis

| Mean | Lower Bound | | 90% CL Mean | | | Upper Bound | Assessment |
|---|---|---|---|---|---|---|---|
| -2.2849 | -5 | > | -10.4592 | 5.8893 | > | 5 | Not equivalent |

| Test | Null | DF | t Value | P-Value |
|---|---|---|---|---|
| Upper | -5 | 79 | 0.55 | 0.2910 |
| Lower | 5 | 79 | -1.48 | 0.0710 |
| Overall | | | | 0.2910 |

*The P value is >0.05 (0.2910) and indicates the methods are not equivalent.*



```
Proc ttest data=iodine tost(-15,15);
paired FR*FFQ;
run;
```

The SAS System

The TTEST Procedure

Difference: FR - FFQ

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 80 | -2.2849 | 43.9283 | 4.9113 | -119.8 | 76.2301 |

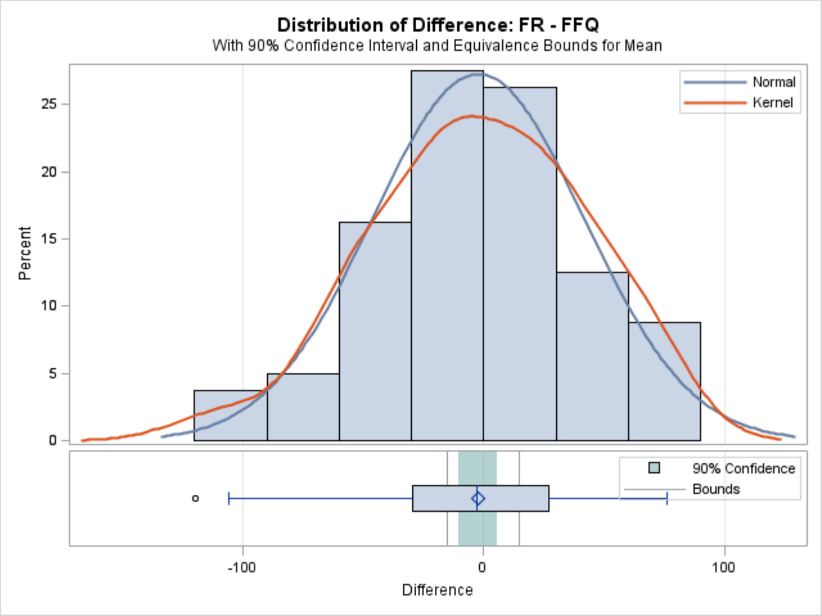
| Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|
| -2.2849 | -12.0607 | 7.4908 | 43.9283 | 38.0178 | 52.0318 |

TOST Level 0.05 Equivalence Analysis

| Mean | Lower Bound | | 90% CL Mean | | | Upper Bound | Assessment |
|---|---|---|---|---|---|---|---|
| -2.2849 | -15 | < | -10.4592 | 5.8893 | < | 15 | Equivalent |

| Test | Null | DF | t Value | P-Value |
|---|---|---|---|---|
| Upper | -15 | 79 | 2.59 | 0.0057 |
| Lower | 15 | 79 | -3.52 | 0.0004 |
| Overall | | | | 0.0057 |

*The overall P value is less than 0.05 and indicates the methods are equivalent*



3. In STATA V12 (StataCorp LP, College Station, TX) equivalence tests are available through a user written .ado file written by Alexis Dinno available from http://doyenne.com/stata/tost.html

```
tostt fr==ffq, eqvt(delta) eqvl(5)

Paired t test of mean equivalence
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      fr |      80    124.2285     5.43549    48.6165    113.4094    135.0476
     ffq |      80    126.5134    6.043544    54.0551     114.484    138.5428
---------+--------------------------------------------------------------------
  D-diff |            7.284939     4.91133               -2.490819     17.0607
  diff+D |            2.715061     4.91133               -7.060696    12.49082
------------------------------------------------------------------------------
mean(diff) =  mean(fr - ffq)
 Delta (D) = 5.0000  Delta expressed in same units as fr

Impossible to reject any Ho if Delta <= t-crit*s.e. ( 8.174 ). See help tostt.

        df = 79


Ho: |diff| >= Delta:

        t1 = 1.483                       t2 = .5528

   Ho1: Delta-diff >= 0         Ho2: diff+Delta <= 0
   Ha1: Delta-diff < 0          Ha2: diff+Delta > 0
   Pr(T > t1) = 0.0710          Pr(T > t2) = 0.2910
```

*Both P values must be significant for the methods to be equivalent, therefore not equivalent*

```
tostt fr==ffq, eqvt(delta) eqvl(15)

Paired t test of mean equivalence
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      fr |      80    124.2285     5.43549     48.6165    113.4094    135.0476
     ffq |      80    126.5134    6.043544     54.0551     114.484    138.5428
---------+--------------------------------------------------------------------
  D-diff |            17.28494     4.91133                7.509181     27.0607
  diff+D |            12.71506     4.91133                2.939304    22.49082
------------------------------------------------------------------------------
mean(diff) =  mean(fr - ffq)
 Delta (D) = 15.0000 Delta expressed in same units as fr
        df = 79

Ho: |diff| >= Delta:

        t1 = 3.519                     t2 = 2.589

   Ho1: Delta-diff >= 0          Ho2: diff+Delta <= 0
   Ha1: Delta-diff < 0           Ha2: diff+Delta > 0
   Pr(T > t1) = 0.0004           Pr(T > t2) = 0.0057
```

*Both P values must be significant for the methods to be equivalent, therefore equivalent*

4. In SPSS V21 (IBM Corporation, Armonk NY) there is not an automated procedure to produce the two one sided tests. This can be done manually by conducting two one sample t tests using the upper and lower equivalence values and the bias as the test variable. This test only produces a two tailed output of significance which needs to be halved for the one tailed P value. If both of these are significant then the methods are equivalent. This can be demonstrated by comparing with the STATA output above. Note that halving the P values is approximate, exact one sided P values could be obtained from several free online calculators or by using R (for example `1-pt(3.519, df=80)` returns P=0.003588676.

```
T-TEST
  /TESTVAL=-5
  /MISSING=ANALYSIS
  /VARIABLES=bias
  /CRITERIA=CI(.95).
T-TEST
  /TESTVAL=5
  /MISSING=ANALYSIS
  /VARIABLES=bias
  /CRITERIA=CI(.95).
```

**One-Sample Test**

| | Test Value = -5 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Mean Difference | Lower | Upper |
| Bias | 1.483 | 79 | .142 | 7.28494 | -2.4908 | 17.0607 |

*The P value of 0.142 must be halved to give 0.071 for the lower equivalence bound*

**One-Sample Test**

| | Test Value = 5 | | | | | |
|---|---|---|---|---|---|---|
| | | | | Mean | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Difference | Lower | Upper |
| Bias | -.553 | 79 | .582 | -2.71506 | -12.4908 | 7.0607 |

*The P value of 0.582 must be halved to give 0.291 for the upper equivalence bound. As neither of these are significant at the 0.05 level, the methods are not equivalent.*

```
T-TEST
  /TESTVAL=-15
  /MISSING=ANALYSIS
  /VARIABLES=bias
  /CRITERIA=CI(.95).
T-TEST
  /TESTVAL=15
  /MISSING=ANALYSIS
  /VARIABLES=bias
  /CRITERIA=CI(.95).
```

**One-Sample Test**

| | Test Value = -15 | | | | | |
|---|---|---|---|---|---|---|
| | | | | Mean | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Difference | Lower | Upper |
| Bias | 3.519 | 79 | .001 | 17.28494 | 7.5092 | 27.0607 |

*The P value of 0.001 must be halved to give P=0.0005 for the lower equivalence bound*

**One-Sample Test**

| | Test Value = 15 | | | | | |
|---|---|---|---|---|---|---|
| | | | | Mean | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Difference | Lower | Upper |
| Bias | -2.589 | 79 | .011 | -12.71506 | -22.4908 | -2.9393 |

*The P value of 0.011 must be halved to give P=0.006 for the upper equivalence bound. As BOTH of these tests are significant at the 0.05 level, the methods are equivalent.*

The SPSS custom dialog box SPSS custom dialog developed by Weber & Popova[50] available from http://www.medianeuroscience.org/equivalence_testing uses effect sizes based on Cohen's d. In order to replicate the examples in this paper the equivalence bounds were converted to approximated effect sizes[51] for the upper and lower bound using the pooled standard deviation of the difference between the methods and correlation from the paired t test and then averaged to create an overall effect size for the 5 and 15 equivalent ranges. The default values for Cohen's small, medium and large effect sizes are also presented as an alternative approach.

*For the equivalence bounds of (-5,5) the approximated effect size is 0.1328*

```
***  Weber & Popova Dependent/Paired-Samples Equivalence Procedure  ***
     Based on the custom-entered delta

                                                    p based on              p based on
                                          actual value of delta   half variance explained
Custom delta          t          df             (two-tailed)            (two-tailed)
____________    ________    ________    _____________________    _______________________

        .133        -.47          79                     .188                       .382
```

*For the equivalence bounds of (-15,15) the approximated effect size is 0.3734*

```
***  Weber & Popova Dependent/Paired-Samples Equivalence Procedure  ***
     Based on the custom-entered delta

                                                    p based on              p based on
                                          actual value of delta   half variance explained
Custom delta          t          df             (two-tailed)            (two-tailed)
____________    ________    ________    _____________________    _______________________

        .373        -.47          79                     .000                       .038
```

Using the default Cohen's effect sizes.

```
***  Weber & Popova Dependent/Paired-Samples Equivalence Procedure  ***
     Based on the Cohen's classification of effect sizes

       t          df        Delta    p, two-tailed
________    ________     ________    _____________

    -.47          79          .10             .291
    -.47          79          .30             .004
    -.47          79          .50             .000
```

References

1. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between 2 methods of clinical measurement. *Lancet* **1**, 307-310.
2. Willett W (1998) *Nutritional Epidemiology*. Second ed, *Monographs in Epidemiology and Biostatistics.* New York: Oxford University Press.
3. Fernandez-Ballart JD, Lluis Pinol J, Zazpe I *et al.* (2010) Relative validity of a semi-quantitative food-frequency questionnaire in an elderly Mediterranean population of Spain. *British Journal of Nutrition* **103**, 1808-1816.
4. Hjartaker A, Andersen LF, Lund E (2007) Comparison of diet measures from a food-frequency questionnaire with measures from repeated 24-hour dietary recalls. The Norwegian Women and Cancer Study. *Public Health Nutrition* **10**, 1094-1103.

5. Government N 8700 Find your ideal figure. www.8700.com.au (accessed 20th November, 2015

6. Council EFI What is the recommended calorie intake for adults, children and toddlers? . http://www.eufic.org/page/en/page/faq/faqid/recommended-calorie-intake-adults-children-toddlers/ (accessed 20th November, 2015

7. Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. *Stat Methods Med Res* **8**, 135-160.

8. Tan L-M, Charlton KE, Tan S-Y *et al.* (2013) Validity and reproducibility of an iodine-specific food frequency questionnaire to estimate dietary iodine intake in older Australians. *Nutrition & Dietetics* **70**, 71-78.

9. Wellek S (2010) *Testing statistical hypothesis of equivalence and noninferiority*. 2nd ed ed. Boca Raton, Fl: Chapman & Hall/CRC.

10. Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm* **15**, 657-680.

11. Westlake WJ (1972) Use of confidence intervals in analysis of comparative bioavailability trials. *J Pharm Sci* **61**, 1340-1341.

12. Midha KK, McKay G (2009) Bioequivalence; Its History, Practice, and Future. *The AAPS Journal* **11**, 664-670.

13. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part 1. *Biometrika* **20**, 175-240.

14. Altman DG, Bland JM (1995) Statistics notes: Absence of evidence is not evidence of absence. *BMJ* **311**, 485.

15. Hoenig JM, Heisey DM (2001) The Abuse of Power. *The American Statistician* **55**, 19-24.

16. Kim Y, Crouter SE, Lee JM *et al.* (2014) Comparisons of prediction equations for estimating energy expenditure in youth. *J Sci Med Sport*.

17. Lange S, Freitag G (2005) Choice of delta: requirements and reality--results of a systematic review. *Biom J* **47**, 12-27; discussion 99-107.

18. Berger RL, Hsu JC (1996) Bioequivalence trials, Intersection-union tests and equivalence confidence sets. *Statistical Science* **11**, 283-319.

19. Groeneveld J (2011) Embedding equivalence t-test results in Bland Altman Plots visualising rater reliability. In *Pharmaceutical Users Software Exchange*, pp. SP06. Brighton, UK: PhUSE.

20. R Core Team (2013) R: A language and environment for statistical computing In *R Foundation for Statistical Computing, Vienna, Austria http://wwwR-projectorg/ ISBN 3-900051-07-0*.

21. Villegas R, Yang G, Liu D *et al.* (2007) Validity and reproducibility of the food-frequency questionnaire used in the Shanghai Men's Health Study. *British Journal of Nutrition* **97**, 993-1000.

22. Senn SS (2008) *Statistical issues in drug development*. 2nd ed, *Statistics in Practice.* Chichester, UK: John Wiley & Sons Ltd.

23. National Health and Medical Research Council Australia and Ministry of Health New Zealand (2nd July, 2014) Nutrient Reference Values for Australia and New Zealand, accessed 12th June 2015. https://www.nrv.gov.au/contact (accessed 12th June, 2015

24. Australian Bureau of Statistics (Last updated 2nd July, 2014) Australian Health Survey: Nutrition First Results - Food and Nutrients, 2011-12. 4364.0.55.007. www.abs.gov.au

25. Charlton KE, Yeatman H, Brock E *et al.* (2013) Improved iodine status in pregnant women 3 years following mandatory iodine fortification of bread in Australia. *Preventative Medicine* **57**, 26-30.

26. Hynes KL, Otahal P, Hay I *et al.* (2013) Mild iodine deficiency during pregnancy is associated with reduced educational outcomes in the offspring: 9-year follow-up of the gestational iodine cohort. *J Clin Endocrinol Metab* **98**, 1954-1962.

27. Weir RR, Carson EL, Mulhern MS *et al.* (2015) Validation of a food frequency questionnaire to determine vitamin D intakes using the method of triads. *Journal of human nutrition and dietetics : the official journal of the British Dietetic Association*.

28. Yuen KK (1974) The two-sample trimmed t for unequal population variances. *Biometrika* **61**, 165-170.

29. Dixon PM, Pechmann JHK (2005) A statistical test to show negligble trend. *Ecology* **86**, 1751-1756.

30. Lombard MJ, Steyn NP, Charlton KE *et al.* (2015) Application and interpretation of multiple statistical tests to evaluate validity of dietary intake assessment methods. *Nutr J* **14**, 40.

31. Reidlinger DP, Willis JM, Whelan K (2015) Resting metabolic rate and anthropometry in older people: a comparison of measured and calculated values. *Journal of Human Nutrition and Dietetics* **28**, 72-84.
32. Siervo M, Bertoli S, Battezzati A *et al.* (2014) Accuracy of predictive equations for the measurement of resting energy expenditure in older subjects. *Clinical Nutrition* **33**, 613-619.
33. Vilar E, Machado A, Garrett A *et al.* (2014) Disease-Specific Predictive Formulas for Energy Expenditure in the Dialysis Population. *Journal of Renal Nutrition* **24**, 243-251.
34. Lazzer S, Patrizi A, De Coi A *et al.* (2014) Prediction of basal metabolic rate in obese children and adolescents considering pubertal stages and anthropometric characteristics or body composition. *European Journal of Clinical Nutrition* **68**, 695-699.
35. Pineau JC, Frey A (2014) Comparison of body composition measurement in highly trained athletes obtained by bioelectrical impedance analysis and dual-energy X-ray absorptiometry. *Science & Sports* **29**, 164-167.
36. Sillanpaa E, Cheng SL, Hakkinen K *et al.* (2014) Body composition in 18-to 88-year-old adultscomparison of multifrequency bioimpedance and dual-energy X-ray absorptiometry. *Obesity* **22**, 101-109.
37. Wan CS, Ward LC, Halim J *et al.* (2014) Bioelectrical impedance analysis to estimate body composition, and change in adiposity, in overweight and obese adolescents: comparison with dual-energy x-ray absorptiometry. *Bmc Pediatrics* **14**.
38. Ziai S, Coriati A, Chabot K *et al.* (2014) Agreement of bioelectric impedance analysis and dual-energy X-ray absorptiometry for body composition evaluation in adults with cystic fibrosis. *Journal of Cystic Fibrosis* **13**, 585-588.
39. Oyeyemi AL, Umar M, Oguche F *et al.* (2014) Accelerometer-determined physical activity and its comparison with the International Physical Activity Questionnaire in a sample of Nigerian adults. *PloS one* **9**, e87233.
40. Aadland E, Ylvisaker E (2015) Reliability of the Actigraph GT3X+ Accelerometer in Adults under Free-Living Conditions. *PLoS One* **10**, e0134606.
41. Vanderloo LM, D'Alimonte NA, Proudfoot NA *et al.* (2015) Comparing the Actical and ActiGraph Approach to Measuring Young Children's Physical Activity Levels and Sedentary Time. *Pediatric exercise science*.
42. Matthys C, Pynaert I, De Keyzer W *et al.* (2007) Validity and reproducibility of an adolescent Web-based food frequency questionnaire. *Journal of the American Dietetic Association* **107**, 605-610.
43. Cullen KW, Watson K, Zakeri I (2008) Relative reliability and validity of the Block Kids Questionnaire among youth aged 10 to 17 years. *Journal of the American Dietetic Association* **108**, 862-866.
44. Brantsaeter AL, Haugen M, Alexander J *et al.* (2008) Validity of a new food frequency questionnaire for pregnant women in the Norwegian Mother and Child Cohort Study (MoBa). *Maternal and Child Nutrition* **4**, 28-43.
45. Toft U, Kristoffersen L, Ladelund S *et al.* (2008) Relative validity of a food frequency questionnaire used in the Inter99 study. *European Journal of Clinical Nutrition* **62**, 1038-1046.
46. Watson JF, Collins CE, Sibbritt DW *et al.* (2009) Reproducibility and comparative validity of a food frequency questionnaire for Australian children and adolescents. *International Journal of Behavioral Nutrition and Physical Activity* **6**.
47. Zhang C-X, Ho SC (2009) Validity and reproducibility of a food frequency questionnaire among Chinese women in Guangdong province. *Asia Pacific Journal of Clinical Nutrition* **18**, 240-250.
48. Ambrosini GL, de Klerk NH, O'Sullivan TA *et al.* (2009) The reliability of a food frequency questionnaire for use among adolescents. *European Journal of Clinical Nutrition* **63**, 1251-1259.
49. Subar AF, Thompson FE, Kipnis V *et al.* (2001) Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires : the Eating at America's Table Study. *Am J Epidemiol* **154**, 1089-1099.
50. Weber R, Popova L (2012) Testing equivalence in communication research: Theory and application. *Communication methods and measures* **6**, 190-213.
51. Dunlap WP, Cortina JM, Vaslow JB *et al.* (1996) Meta-analysis of experiments with matched groups of repeated measures designs. *Psychological Methods* **1**, 170-177.