

Reliability and validity of food portion size estimation from images using manual flexible digital virtual meshes

Alicia Beltran¹, Hafza Dadabhoj¹, Courtney Ryan¹, Ruchita Dholakia¹, Janice Baranowski¹, Yuecheng Li², Guifang Yan³, Wenyan Jia², Minguo Sun² and Tom Baranowski^{1,*}

¹USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, 1100 Bates Street, Houston, TX 77030, USA; ²Department of Neurological Surgery, University of Pittsburgh, Pittsburgh, PA, USA; ³Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA

Submitted 6 October 2017: Final revision received 14 December 2017: Accepted 21 December 2017: First published online 12 February 2018

Abstract

Objective: The eButton takes frontal images at 4 s intervals throughout the day. A three-dimensional manually administered wire mesh procedure has been developed to quantify portion sizes from the two-dimensional images. The present paper reports a test of the inter-rater reliability and validity of use of the wire mesh procedure.

Design: Seventeen foods of diverse shapes and sizes served on plates, bowls and cups were selected to rigorously test the portion assessment procedure. A dietitian not involved in inter-rater reliability assessment used standard cups to independently measure the quantities of foods to generate the 'true' value for a total of seventy-five 'served' and seventy-five smaller 'left' images with diverse portion sizes.

Setting: The images appeared on the computer to which the digital wire meshes were applied.

Subjects: Two dietitians and three engineers independently estimated portion size of the larger ('served') and smaller ('left') images for the same foods.

Results: The engineers had higher reliability and validity than the dietitians. The dietitians had lower reliabilities and validities for the smaller more irregular images, but the engineers did not, suggesting training could overcome this limitation. The lower reliabilities and validities for foods served in bowls, compared with plates, suggest difficulties with the curved nature of the bowls.

Conclusions: The wire mesh procedure is an important step forward in quantifying portion size, which has been subject to substantial self-report error. Improved training procedures are needed to overcome the identified problems.

Keywords
Wire mesh
Portion size
Images
Reliability
Validity
Diet

Assessing portion sizes of dietary intake with high accuracy (validity) has been a challenge among adults⁽¹⁾ and children⁽²⁾. The most accurate method for estimating total energy intake is the 24 h dietary recall⁽³⁾. However, substantial self-reporting error remains in this method⁽⁴⁾.

To minimize errors of self-report, novel technologies have used images of foods to improve the accuracy of food identification and portion size estimation⁽⁵⁾. While different image-based methods estimate portion size in different ways, the active methods require placement of a fiduciary marker in the smartphone camera's field of view⁽⁶⁾. The eButton, a passive recording system which takes images of everything in front of a person at 4 s intervals throughout the day⁽⁷⁾, uses three-dimensional flexible digital wire meshes that can be superimposed on

the two-dimensional food image, manually distended or manipulated to closely adhere to the outer edge of the food image, and the volume of the three-dimensional image used to estimate the food portion size⁽⁸⁾.

The accuracy of this eButton three-dimensional wire mesh portion size estimation procedure has been tested with 100 Asian and Western food samples⁽⁸⁾. After selecting the images with food events before and after a meal, the most appropriate shaped wire mesh was selected from a virtual wire mesh shape library, and used to estimate volume and calculate portion size⁽⁸⁾. The wire mesh procedure employed by its designing engineers with food images obtained under naturalistic circumstances resulted in a mean relative error of 2.8% (SD 20.4%) against an objective index of volume (seed displacement). This mean relative error was substantially

*Corresponding author: Email tbaranow@bcm.edu

better than three human visual raters of portion size, whether they knew the sizes of the plates or did not. Our previous research with a small sample indicated that engineers scored higher validity than dietitians⁽⁷⁾. Since our ultimate goal is to automate use of the wire mesh, we wanted to estimate its most accurate possible use.

The present study tested the precision (reliability) and accuracy (validity) of the use of three-dimensional wire mesh shapes as used by two experienced dietitians and three engineers from the lab that created the procedure. Validity was assessed against a known true value (but unknown by the dietitians or engineers) with the same foods and dishes, but with different portion sizes, as if before ('served') and after ('left') a meal. We expected high reliability between the two trained dietitians for the 'served' images (comparable to that of the engineers), since the portions were reasonably large and in usual portions, but somewhat lower reliability for the smaller irregular 'left' images since we believed the smaller images would present more problems in precisely deforming the wire meshes. The 'intake' (calculated from 'served' minus 'left') quantity (not an image) reliability was expected to be even lower, since these values reflect any unreliability in quantifying the 'served' and 'left' images. We expected the validity intraclass correlation coefficients (ICC; including a known 'true' value) to be only slightly lower than the reliability ICC since the previous report had 2.8% relative error.

Methods

Seventeen foods of diverse shapes and sizes served in different plates, bowls and cups were selected (Table 1) to rigorously test the portion assessment procedure. Since the wire mesh procedure specifies volume, we elected to measure the criterion using a unit of volume (i.e. cups). A third dietitian carefully used standard cups

to independently measure the quantities of food to generate the 'true' value for a total of seventy-five 'served' and seventy-five smaller 'left' intake images with diverse different portion sizes. For apple, orange, cheesecake wedge, slice of cake and pizza slice (those not amenable to cup assessment), the dietitian weighed the quantity to obtain the portion size, which was converted to cup equivalents using the food density obtained from the Food and Nutrient Database of Dietary Studies (FNDDS)⁽⁹⁾ database incorporated into the eButton. For each food, images were taken in different types of dinnerware (i.e. large bowl, small bowl and plain plate for food items and glass and plastic cup for drinks; Table 1). The images were presented to the dietitians and the engineers in the sequence specified by the dietitian who created this testing activity. This sequence was not randomized, but there was no attempt to simplify or complicate the sequence of images.





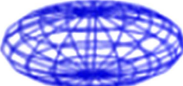




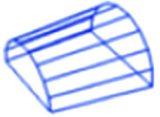
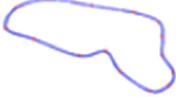
The eButton library of wire meshes contains eleven different shapes common to foods, including cuboid, cylinder, sphere, wedge, ellipse, half sphere, half ellipse, half sphere, section of sphere, tunnel and an irregular shape (Table 2). Based on the food captured in the image, the dietitians and engineers selected (independently) the wire mesh shape that best fit the food (Table 1) and the wire mesh was virtually fitted (using digital pressure points) as tightly as possible to the outer perimeter of the food image to estimate volume. Then a computer algorithm estimated the volume of the fitted wire frame to estimate portion size⁽⁸⁾. This procedure was followed for both 'served' and 'left' images.

Reference points are needed to calculate the real size of the food in the image. The diameter and depth of standard plates, bowls and glasses (the same as used in the images) were measured and entered into the software to calculate the food's dimensions in relation with the image. With the dinnerware measurements, the software prompts to select six reference points, which are manually added to the image around the border of the circular dinnerware.

Table 1 Foods selected for portion size validations with different dishes and wire frames

Food	Dishes	Wire meshes used
1. Scrambled egg	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
2. Long-grain brown rice	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
3. Old fashioned oatmeal	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
4. Whole kernel corn	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
5. Organic romaine lettuce	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
6. No-salt-added black beans, canned	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
7. Fruit cocktail, canned	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
8. Mashed sweet potatoes	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
9. Frozen broccoli	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
10. Campbell's vegetable soup	Large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere
11. H-E-B Half and Half Instant Tea	Glass, plastic cup	Cylinder
12. Apple	Plate, small bowl	Before picture: half ellipse, half sphere, section of sphere, sphere After picture: half ellipse, half sphere, section of sphere
13. Cheesecake	Plate	Wedge, half ellipse
14. Orange	Plate, small bowl	Before picture: half ellipse, sphere After picture: half ellipse, section of sphere
15. Pizza slice	Plate	Wedge, half ellipse
16. Cake	Plate	Cuboid, half ellipse, cylinder
17. Kraft Macaroni and Cheese Spirals	Plate, large bowl, small bowl	Half ellipse, half sphere, cylinder, section of sphere

Table 2 Wire frames library for portion size volume estimation

Wire frame model	Appearance	Examples
Cuboid		Cornbread, cake
Cylinder		Drinks, liquids in bowls & glasses
Sphere		Peach, apple, orange
Wedge		Pizza, pie, cake
Ellipse		Potato
Half sphere		Half orange, face down
Half ellipse		Mounds of food (pasta, rice, stews)
Half sphere		Half peach, face up; large mounds
Section of sphere		Mounds of food
Tunnel		Cut carrot
Irregular shape		Steak, chicken

The volume of the wire mesh (and thereby the food) can be estimated with the correspondence of the spatial points and the picture points⁽⁸⁾.

Two dietitians (not including the third dietitian who created the test images) with extensive experience in use of the wire mesh procedure in two earlier studies⁽⁷⁾ (but no

specialized training) and three engineering students who participated in the programming of the eButton software (but without any training in use of the eButton wire mesh procedure) independently estimated portion size of the larger ('served') and smaller ('left') images for the same foods. The wire mesh volumes were estimated by each rater

independently, blinded from both the true measure and each other's estimates.

Data analysis

The primary output of the wire mesh is 'volume' in cubic centimetres. The reliability-related ICC⁽¹⁰⁾ was used between the two dietitians to assess reliability of portion assessment, and across the two dietitians and the true value to assess validity of portion assessment. The equation for ICC is:

$$ICC = (\text{no. of raters} / \text{no. of raters} - 1) \times (1 - \text{individual variance} / \text{sum of variances}).$$

The volume assessments were provided by the eButton software using FNDDS 2013–2014 data⁽¹¹⁾ for the volume of the food in the images. There were 150 total images (including seventy-five 'served' and seventy-five 'left' images) across all foods and dinnerware. Table 3 includes the ICC and 95 % CI for reliability and validity for the dietitians and engineers separately. Table 4 presents the same coefficients and 95 % CI, but for subsets by type of dinnerware. Since separating the images into serving container type

categories resulted in a small number of images per category, the served and left images were combined for these analyses alone. Table 5 presents estimates of relative error percentages (true value – estimated value/true value) for dietitians and engineers separately. Figure 1 presents box-and-whisker plots of the relative error percentages to depict the distributions of error by type of image and professionals.

Results

Inter-rater (between the two dietitians) reliability ICC were acceptably high for volume served (0.771, 95 % CI 0.660, 0.849), but lower for volume left (0.629, 95 % CI 0.470, 0.748) and intake volume (0.590, 95 % CI 0.420, 0.720). When the ICC were calculated to include the true value (i.e. the amount actually included in the images by the third dietitian), the validity ICC were similar for volume served (0.766, 95 % CI 0.680, 0.836) and volume left (0.596, 95 % CI 0.474, 0.706), but increased substantially for intake volume (0.677, 95 % CI 0.569, 0.769; Table 3). The reliability coefficients for the three engineers were higher than for the dietitians for volume

Table 3 Reliability coefficients with 95 % CI for different portion variables across different raters and the true value

Variable	Images <i>n</i>	Inter-rater reliability (for two dietitians)		Validity (for two dietitians+ true value)		Inter-rater reliability (for three engineers)		Validity (for three engineers + true value)	
		ICC	95 % CI	ICC	95 % CI	ICC	95 % CI	ICC	95 % CI
Volume served	75	0.771	0.660, 0.849	0.766	0.680, 0.836	0.820	0.750, 0.876	0.815	0.750, 0.869
Volume left	75	0.629	0.470, 0.748	0.596	0.474, 0.706	0.847	0.786, 0.895	0.832	0.772, 0.881
Intake volume	75	0.590	0.420, 0.720	0.677	0.569, 0.769	0.755	0.666, 0.828	0.745	0.664, 0.817

ICC, intraclass correlation.

Intake is calculated as the difference between the initial 'serving' and what is 'left' on the plate, but if the estimated volume of the leftover is larger than the served, the intake is set to 0.

Table 4 Intraclass correlation coefficients (ICC) with 95 % CI for different serving containers

Volume ('served' and 'left' images combined)	Small bowl (<i>n</i> 48)		Large bowl (<i>n</i> 42)		Plate (<i>n</i> 56)	
	ICC	95 % CI	ICC	95 % CI	ICC	95 % CI
For two dietitians	0.669	0.529, 0.784	0.712	0.574, 0.821	0.833	0.755, 0.892
For three engineers	0.820	0.739, 0.885	0.844	0.767, 0.904	0.877	0.823, 0.919

The total number of images in this analysis is 146. Four 'glass' images were excluded since the same was too small to reliably estimate the ICC. Since separating the images into serving container type categories resulted in a small number of images per category, the 'served' and 'left' images were combined for these analyses alone.

Table 5 Statistics of relative error percentages for dietitians and engineers separately

Measure	'Served'			'Left'			'Intake'		
	Images (<i>n</i>)	Median (%)	IQR (%)	Images (<i>n</i>)	Median (%)	IQR (%)	Images (<i>n</i>)*	Median (%)	IQR (%)
For all dietitians	150	17.9	58.7	150	50.7	130.2	148	-4.6	69.2
For all engineers	225	3.7	51.4	225	14.1	62.3	222	-0.03	63.2

Relative error = [(observed – true)/true] × 100; IQR, interquartile range (also called the mid spread or middle 50 %, equals the difference between the 75th and 25th percentiles).

Note: (i) *Since the real food intake in one image is zero, the relative errors from two dietitians and three engineers for this food had to be eliminated. (ii) For median values of relative error: intake < served < left. (iii) Big estimation variance can be seen in IQR, especially for 'left' food, probably due to the small size of the leftover. (iv) Median values ('served' and 'left') for dietitians are positive and larger than the estimates from engineers. This may be caused by the fitting procedure if outer perimeter is used. But fortunately, when calculating the volume of intake, such bias can be cancelled to some extent. (v) Overestimation can be seen in all of these conditions.

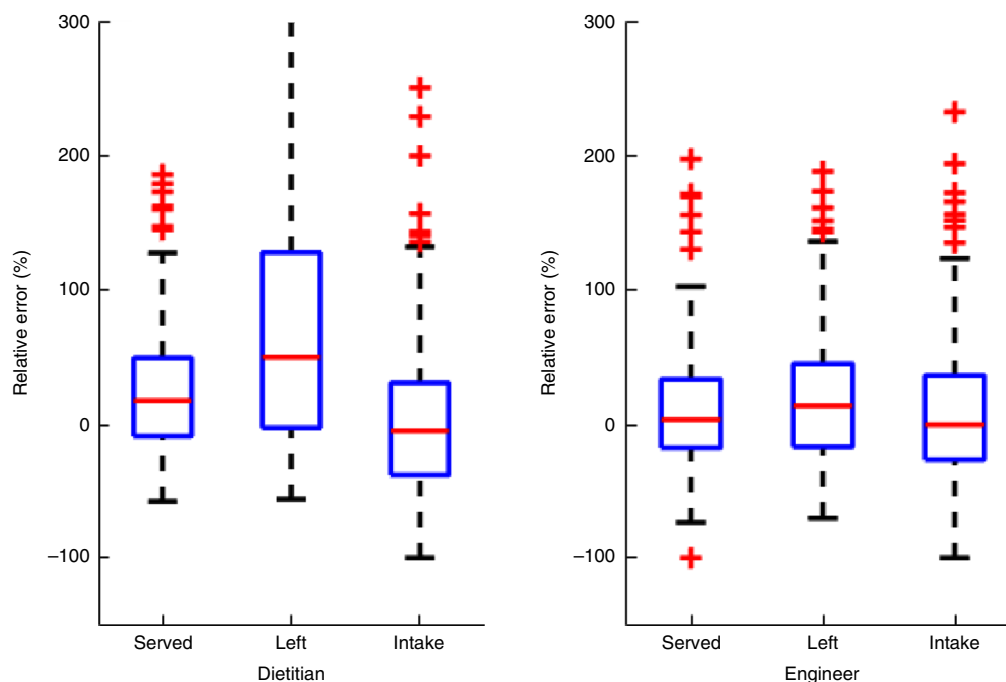


Fig. 1 (colour online) Box-and-whisker plot of relative error percentages for the two dietitians and three engineers. On each box, the central line represents the median of the relative errors over all the food samples. The bottom and top edges of the box are respectively the first and third quartiles, which is the interquartile range (IQR). The extreme regions (with a greater distance from the median than 1.5 times the IQR) are the ends of the lines extending from the IQR. Points outside this region are plotted individually as plus signs, representing potential outliers. The highest relative error for dietitians for the 'left' images was 1100, which was substantially off the scale, the inclusion of which distorted the scale representation for all the other cases

served (0.820, 95% CI 0.750, 0.876), volume left (0.847, 95% CI 0.786, 0.895) and intake volume (0.755, 95% CI 0.666, 0.828; Table 3). The same was true for their validity coefficients.

When the same reliability ICC for intake volumes were estimated by type of serving container, the highest ICC were obtained for the plate (Table 4). The median relative error percentage across the dietitians for volume was 17.9% for served, 50.7% for left and -4.6% for intake (Table 5). The engineers obtained substantially lower median relative error percentages for volume, with 3.7% for served, 14.1% for left and -0.03% for intake. A box-and-whisker plot of the data in Table 5 (see Fig. 1) demonstrates the distributional differences in the relative error percentages between the dietitians and engineers.

Discussion

We anticipated high reliability and validity by both the dietitians and the engineers for the larger portions in the 'served' images, but lower reliability for the smaller 'left' images, which were more irregular and thereby harder to precisely fit with a wire mesh. As expected, the inter-recorder reliability (r) and validity (v) for volume served ($r=0.771$, $v=0.766$) by experienced dietitians from food images was reasonably high, but lower for volume left ($r=0.629$, $v=0.596$; Table 3). One immediately notices

that both the reliability and validity values obtained by the engineers were higher for volume served ($r=0.820$, $v=0.815$) and substantially higher for volume left ($r=0.847$, $v=0.832$). Lower reliability was expected for the calculated intake volume, since the 'serving' minus 'left' values would combine errors from the two estimates. This was true for both the dietitians ($r=0.590$, $v=0.677$) and the engineers ($r=0.755$, $v=0.745$). Furthermore, reliability and validity values were consistently high among the engineers, whereas the validity values were lower for the dietitians than the engineers. The relative error percentages (Table 5) support this pattern of findings, suggesting that the dietitians had not mastered the manipulation of the wire mesh to closely conform to the outer boundary of the food image whereas the engineers who helped create the wire imaging system did. While the lower reliability and validity for intake volume reflects the unreliability in both the volumes served and left, and thereby was expected, this indicates substantial error in using the difference to estimate intake (even for the engineers).

The validity ICC for volume served for the dietitians (0.766) and the engineers (0.815) were comparable to the reliability ICC for volume served (0.771, 0.820), reflecting reasonable concordance between raters with the true value. The substantial drop in the validity ICC for volume left for the dietitians (0.596), but not for the engineers (0.832), suggests that dietitians had difficulty estimating

the smaller irregular amounts, which might be improved, perhaps through training, to approximate that of the engineers.

A review of the images revealed several problems among the dietitians and engineers (but more among the dietitians): (i) when a portion of the plate or bowl was not fully within the image, the mesh could not fit the food completely; (ii) the mesh was not always adjusted to completely and closely fit the food image; (iii) when the shape of the food was irregular, the wire mesh could not precisely fit the food image; (iv) the size of the plate or bowl was occasionally mis-specified; and (v) some foods in FNDDS did not have density, thereby introducing inaccuracies in conversions of weight to volume and vice versa. Enhanced training will minimize some of these errors, but further technological development is needed to address issue (iii) and improved food analytic data are needed to address issue (v).

The reliability ICC for the three serving containers were higher for the plates than for the bowls for both the dietitians and the engineers, and somewhat lower for the smaller than the larger bowls (Table 4). The likely explanation is it was more difficult to insert the wire meshes below the line of vision into the bottom of the opaque bowls, and the curved nature of the bowls, thereby leading to more error.

The engineers' use of the wire meshes resulted in consistently higher reliability and validity ICC and substantially lower median relative error percentages than the dietitians. Furthermore, the engineers' validity coefficients (ICC across the engineers and true value) were almost as large as their reliability coefficients (Table 3), suggesting high accuracy in use of the wire mesh method is possible. The median relative error percentage was not as low as that reported in a previous paper from their group (2.8%)⁽⁸⁾, suggesting these engineering students could improve their technique, especially for the smaller portions of food 'left' (14.1%). The substantially lower median reliabilities and validities and the higher median error for dietitians indicates that the engineers knew better how to correct for problems/limitations in the wire mesh procedure. Despite their substantial prior experience in using the meshes, the dietitians could benefit from more thorough training including trial-by-trial feedback on their performance.

Desirable training in the use of the meshes for credentialed dietitians (i.e. who know the intricacies of 24 h recall methods) without previous experience with the passive method would likely include: a video demonstration of the use of the multiple flex points per wire mesh to deform or reform the wire mesh; application of the different wire meshes to diverse images of foods (e.g. mounds, pie, irregular pieces of meat, soup in a bowl); demonstration of common errors in using the wire mesh (e.g. wire mesh boundary not tightly conforming to the food image boundary, tubular wire mesh not distending to conform to the curved nature of a bowl); practice in using the wire

mesh with diverse foods; and feedback on the practice volume estimations. At the end of the training, trainees should be tested on images and allowed to retake the test as needed to pass. The number of images and the criterion for acceptable accuracy need further research to accurately determine performance in real-world circumstances.

A limitation of the current study is that all eleven meshes were not adequately tested. Additional research will be needed employing a number of foods for which each of the underused meshes is appropriate to estimate the same values. The dietitians and engineers were allowed to select the mesh they thought most appropriate to each food, with substantial differences in which wire meshes were used per food. Part of the error may reflect differences in appropriateness of each wire mesh for each food. More thorough analysis needs to be done on errors that the dietitians made, but were not by the engineers, to identify ways to enhance future dietitian performance.

Conclusion

The wire mesh procedure for estimating portion size in food images is an important step forward in quantifying a variable which has been subject to substantial self-report error. The engineers who helped create the procedure had higher reliability and validity than the dietitians who had substantial prior experience in using the procedure, but no direct feedback on the accuracy of their estimates, suggesting more intensive analysis is needed to identify where the biggest errors occurred and training procedures are needed that can overcome them. Additional development with the meshes and training in use of the meshes are needed to advance this method of dietary assessment.

Acknowledgements

Financial support: This research was funded by a grant from the National Institutes of Health, National Cancer Institute (number 5 R21 CA172864) and institutional support from the US Department of Agriculture, Agricultural Research Service (Cooperative Agreement 58-3092-5-001). The funders had no role in the design, analysis or writing of this article. *Conflict of interest:* The authors have no disclosures to report. *Authorship:* A.B. conceptualized the study, was the project manager, conducted data collection and use of the wire mesh, and wrote a preliminary draft of the manuscript. H.D. participated in data collection and use of the wire mesh. C.R. created the portion images. R.D. conducted the statistical analyses. J.B. was the project coordinator. Y.L. and G.L. helped create the wire meshes and participated in use of the wire meshes on the images. W.J. helped create the wire meshes, managed the engineer use of wire meshes and conducted statistical analyses. M.S. was principal investigator for creation of the eButton and conceived of the wire meshes. T.B. was principal

investigator for the overall project, helped conceptualize this study and wrote the successive drafts of the manuscript. All authors reviewed, critiqued and approved this manuscript. *Ethics of human subject participation*: Not applicable.

References

1. Tabacchi G, Amodio E, Di Pasquale M *et al.* (2014) Validation and reproducibility of dietary assessment methods in adolescents: a systematic literature review. *Public Health Nutr* **17**, 2700–2714.
2. Foster E, Adamson AJ, Anderson AS *et al.* (2009) Estimation of portion size in children's dietary assessment: lessons learnt. *Eur J Clin Nutr* **63**, Suppl. 1, S45–S49.
3. Burrows TL, Martin RJ & Collins CE (2010) A systematic review of the validity of dietary assessment methods in children when compared with the method of doubly labeled water. *J Am Diet Assoc* **110**, 1501–1510.
4. Archer E, Pavea G & Lavie CJ (2015) The inadmissibility of What We Eat in America and NHANES dietary data in nutrition and obesity research and the scientific formulation of national dietary guidelines. *Mayo Clin Proc* **90**, 911–926.
5. Sun M, Fernstrom JD, Jia W *et al.* (2010) A wearable electronic system for objective dietary assessment. *J Am Diet Assoc* **110**, 45–47.
6. Fang S, Liu C, Zhu F *et al.* (2015) Single-view food portion estimation based on geometric models. *ISM* **2015**, 385–390.
7. Beltran A, Dadabhoy H, Chen TA *et al.* (2016) Adapting the eButton to the abilities of children for diet assessment. In *Proceedings of Measuring Behavior 2016 – 10th International Conference on Methods and Techniques in Behavioral Research*, pp. 72–81 [A Spink, G Riedel, L Zhou *et al.*, editors]. http://www.measuringbehavior.org/files/2016/MB2016_Proceedings.pdf (accessed January 2018).
8. Jia W, Chen HC, Yue Y *et al.* (2014) Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. *Public Health Nutr* **17**, 1671–1681.
9. US Department of Agriculture, Agricultural Research Service (2012) USDA Food and Nutrient Database for Dietary Studies, 5.0. <http://www.ars.usda.gov/ba/bhnrc/fsrg> (accessed January 2018).
10. Shrout PE & Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* **86**, 420–428.
11. US Department of Agriculture, Agricultural Research Service (2016) USDA Food and Nutrient Database for Dietary Studies 2013–2014. <http://www.ars.usda.gov/nea/bhnrc/fsrg> (accessed January 2018).