Evaluation of a Coding System
for a Smartphone Based Visual Food Record

Sabine Arnold
University of Konstanz
Psychological Assessment and Health Psychology

A thesis submitted in partial fulfillment of the requirements of the degree
Master of Science (M.Sc.)

Primary Supervisor: Professor Dr. Britta Renner
Secondary Supervisor: Professor Dr. Harald Schupp

October 23rd, 2015
# Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of contents</td>
<td>I</td>
</tr>
<tr>
<td>List of figures</td>
<td>III</td>
</tr>
<tr>
<td>List of tables</td>
<td>IV</td>
</tr>
<tr>
<td>Abstract</td>
<td>V</td>
</tr>
</tbody>
</table>

## 1. Introduction

1.1 Self-reported dietary assessment methods                               | 1
   1.1.1 Food frequency questionnaire                                       | 2
   1.1.2 24-hour recall                                                    | 2
   1.1.3 Food photographs                                                  | 4
   1.1.4 Self-reported dietary assessment methods: Summary                 | 7
1.2 A visual food record: Assessment and coding methods                   | 10
   1.2.1 A visual food record based on a self-developed device            | 12
   1.2.2 A smartphone based visual food record                            | 12
   1.2.3 A Visual food record: Summary                                    | 15
1.3 Present study                                                        | 21
   1.3.1 Summary of the current research                                  | 24
   1.3.2 Present food record and coding system                            | 26
   1.3.3 Research questions                                                | 27

## 2. Methods

2.1 Participants                                                         | 28
2.2 Procedure                                                           | 28
2.3 Coding system                                                       | 29
   2.3.1 Coding procedure                                                | 29
   2.3.2 Food atlas                                                       | 31
   2.3.3 OptiDiet                                                        | 31
2.4 Data analysis                                                       | 32
   2.4.1 Krippendorff’s alpha                                             | 32
   2.4.2 Excluded cases                                                  | 33
   2.4.3 Overview                                                        | 34

## 3. Results

3.1 How much people ate: Food weight estimation                           | 34
   3.1.1 Food weight estimation for all participants                      | 34
   3.1.2 Food weight estimation per participant                          | 36
3.2 What people ate: Food labeling                                       | 37
   3.2.1 Number of assigned food labels for all participants             | 37
   3.2.2 Number of assigned food labels per participant                 | 38
   3.2.3 Content of assigned food labels                                 | 39
3.3 What people’s food consisted of: Food Nutrients                      | 40
   3.3.1 Reliability for all nutrients                                   | 40
   3.3.2 Reliability per nutrient                                        | 40
4. Discussion

4.1 The coding system’s reliability for food weight

4.1.1 Reliability for food weight for all and per participant

4.1.2 Future research on the coding system’s reliability for food weight

4.2 The coding system’s reliability for food labels

4.2.1 Reliability for food labels for all and per participant

4.2.2 Future research on the coding system’s reliability for food labels

4.3 The coding system’s reliability for food nutrients

4.3.1 Reliability for all and per nutrient

4.3.2 Future research on the coding system’s reliability for food nutrients

4.4 Limitations

4.4.1 Excluded cases

4.4.2 Participants

4.4.3 Coders

4.4.4 Kalpha

4.5 Future research: Validation of the coding system

4.6 Conclusion

References

Appendix
List of figures

Figure 1 Technology Assisted Dietary Assessment System (TADA) (Boushey et al., 2015) . . . 18
Figure 2 Procedure of the smartphone based visual food record (Renner et al., 2015) . . . . . 30
Figure 3 Example of weight gradation in the food atlas ................................................. 31
Figure 4 Estimated weight per meal by coders A and B for all participants ......................... 35
Figure 5 Percentage of meals per difference in estimated weight between coders A and B for all participants ................................................................. 36
Figure 6 Estimated weight per meal by coders A and B per participant ................................ 38
Figure 7 Percentage of meals per difference in estimated weight between coders A and B for each participant ................................................................. 39
Figure 8 Number of assigned labels per meal by coders A and B for all participants .......... 41
Figure 9 Percentage of meals per difference in the number of assigned labels between coders A and B ................................................................. 42
Figure 10 Number of assigned labels per meal by coders A and B per participant ............... 44
Figure 11 Percentage of meals per difference in the number of assigned labels between coders A and B per participant ................................................................. 46
Figure 12 Percentage of food items assigned per food category by coders A and B .......... 48
Figure 13 Kalpha for micro- and macronutrients according to the GFFS ......................... 51
Figure 14 Extension of the food atlas ................................................................................. 57
List of tables

Table 1 Characteristics of self-reported dietary assessment methods ......................... 11
Table 2 Characteristics of visual food records .......................................................... 22
Table 3 T-test for the food weight estimation ......................................................... 40
Table 4 Differences in the food weight estimation by percentiles ............................... 78
Table 5 T-test for the number of assigned labels ..................................................... 45
Table 6 Differences in the number of assigned labels ............................................... 47
Table 7 Food items per GFFS food category ............................................................. 49
Table 8 Kalpa for nutrients ....................................................................................... 52
Table 9 Limitations and questions for future research ............................................. 69
Table 10 Overview of reliabilities and ideas for their improvements ......................... 70
Abstract

Among available dietary assessment tools the smartphone stands out in up-to-date measurement but requires further development. This study used data from Renner et al.’s (2015) SmartFood study to test the reliability of the smartphone based visual food record Renner et al. (2015) developed for Germany. In the SmartFood study participants photographed their diet using a smartphone within everyday living conditions over eight days. To determine the coding system’s reliability, two coders independently coded all 269 digital food images from 10 participants (six males, four females; age 30–74 years). Renner et al.’s (2015) coding system, comprising a food atlas (digital booklet with food images), and the nutritional software OptiDiet© were used to analyze the visual food record. As intercoder reliability index Krippendorff’s alpha was calculated to measure coder agreement. Across all participants’ meals, reliability was medium to high with $\alpha = .71$, $SD = 0.17$, 95% CI [.59, .84] for food weight, $\alpha = .76$, $SD = 0.11$, 95% CI [.69, .84] for food labeling, and $\alpha = .67$, $SD = 0.08$, 95% CI [.59, .72] for food nutrients. Reliability varied in dependence of participant and nutrient type but was not significantly affected by the coders. The results support the coding system as a promising tool for analyzing smartphone based visual food records. To further develop the coding system, the food atlas should be extended with more images and the coding guidelines revised. Finally, a validation study is required.

Key words: visual food record, smartphone, food atlas, coding, intercoder reliability, food weight, food labeling, nutrient
Introduction

How much and what kinds of food do people eat in their everyday lives? Reliable information on this issue is requisite for several fields of research and practice. Specifically, the amount of food consumed, food type, and food nutrients constitute key factors in dietary focused studies in sports (e.g., Loucks, Kiens, & Wright, 2011), medicine (e.g., Livingstone & Pourshahidi, 2014), and psychology (e.g., Meule, von Rezori, & Blechert, 2014). In terms of nutritional diseases such as obesity (Livingstone & Pourshahidi, 2014) or bulimia nervosa (Meule et al., 2014), knowledge of a person’s daily food intake is crucial for designing effective medical and psychotherapeutic interventions. Furthermore, the study of a positive and healthy eating style in non-clinical populations requires reliable data on how much and what people eat.

Today, various methods are available for recording food (e.g., Illner, Freisling, Boeing, Huybrechts, Crispim, & Slimani, 2012; Shim, Oh, & Kim, 2014). To compare them, the present study considered the following criteria: First, to obtain comprehensive information, the food record should allow for assessment of both food weight (how much people ate) and food type (what people ate). Second, to gain insight into a person’s actual food consumption, data collection should also be possible under real life conditions and not only in the laboratory. Third, the method should be convenient for the participant. Fourth, and most importantly, to ensure that the food record measures precisely what it claims to measure, it must show sound psychometric quality. To this end, reliability and validity must be tested.

Considering the criteria listed above, the present study begins with a summary of self-reported dietary assessment methods. The most outstanding up-to-date measuring system, the visual food record, is then analyzed in depth. Here, several approaches for assessing and coding a visual food record are introduced. Moreover, a newly developed coding system for a smartphone
Introduction

based visual food record for Germany (Renner, Sproesser, Koenig, Ziesemer, & Schupp, 2015) is presented. The present study determines the reliability of this coding system precisely to obtain a first index of its psychometric quality.

1.1 Self-reported dietary assessment methods

Among the numerous dietary assessment methods available the food frequency questionnaire (FFQ) and the 24-hour recall (24-HR) are two of the most common subjective measures used (Magarey et al., 2011; Shim et al., 2014). Both require participants to report their food intake themselves and use food photographs (FP) (e.g., Brito, Guimarães, & Pereira, 2014; Lazarte, Encinas, Alegre, & Granfeldt, 2012)

1.1.1 Food frequency questionnaire

As a self-reported dietary assessment method, the FFQ has a long history in epidemiological studies (e.g., Shim et al., 2014). It relies on the frequency of consumed food to assess a participant’s usual dietary intake (Rusin, Årsand, & Hartvigsen, 2013). As described by Kolodziejczyk, Merchant, and Norman (2012) a food list is presented, and the participant is asked to indicate how often he or she consumed the respective foods during a certain period. Usually a food list includes 100 to 150 foods (Shim et al., 2014). While some FFQs focus on frequency, others also record portion size (Wakai, 2009). Hence, a FFQ is capable of assessing both how much and what people ate. Moreover, the FFQ can be interviewer- or self-administered (Shim et al., 2014).

Although the FFQ was originally a paper-and-pencil test, computer- and web-based formats are now available (Shim et al., 2014). For example, in an interactive computer-based version, an audio script can read aloud reminders to fill in the FFQ (Wong, Boushey, Novotny, & Gustafson, 2008).
Generally, the FFQ has been applied in large epidemiological studies (Shim et al., 2014). Specifically, it established a principal research tool in nutrition epidemiology. According to Cade, Burles, Warm, Thompson, and Margetts (2004), the FFQ was designed to analyze a specific disease in 56.71% (93/164) of the investigated studies.

Regarding the FFQ’s reliability, some exemplary recent evidence is available. Retest reliability is often assessed using two measurement times (e.g., Fernández-Ballart et al., 2010; Jackson, Walker, Younger, & Bennett, 2011; Wakai, 2009). Single studies revealed a retest reliability ranging from $r = .50$ to $.82$ (Fernández-Ballart et al., 2010) or $r = .50$ to $.84$ (Jackson et al., 2011). Findings from Kolodziejczyk et al.’s (2012) review also showed the FFQ’s reliability in a range between $r = .05$ and $.88$, and Wakai (2009) reviewed the FFQ’s reliability as varying between $r = .50$ and $.72$ (median correlation).

To investigate validity, a recent study by McGowan, Curran, and McAuliffe (2014) compared participants’ ranking of nutrient intake recorded in FFQ and food diaries and found a validity ranging from $r = .24$ to $.59$. Another study (Fernández-Ballart et al., 2010) drew a comparison between the FFQ and food diaries by taking nutrients, energy and food groups as dependent variables. Here, the FFQ’s validity varied between $r = .24$ and $.72$. Moreover, by comparing a FFQ with a 24-HR recall, Jackson et al. (2011) found a correlation ranging from $r = .38$ to $.86$ for nutrients. Summarized by Kolodziejczyk et al. (2012), the FFQ’s validity data ranged from $r = .01$ to $.80$, and Wakai (2009) found coefficients to be between $r = .31$ and $.56$ (median correlation).

To further describe the FFQ, some of its advantages are outlined here. For instance, the FFQ imposes a low respondent burden (Illner et al., 2012) and can be administered with ease (Cade et al., 2004). Other advantages are its cost- (Illner et al., 2012; Shim et al., 2014) and time-effectiveness (Shim et al., 2014), which can probably be attributed to its attempt to estimate...
multiple 24-hour recalls, data analysis for the FFQ is less time intensive and less expensive (Kolodziejczyk et al., 2012). Furthermore, modern FFQ versions offer additional advantages. In particular, web-based FFQs can be completed at any time and location, making them easy for geographically spread samples to access (Illner et al., 2012).

Aside from these benefits, the FFQ also contains limitations. For example, since the FFQ relies on memory (Burrows, Martin, & Collins, 2010), a participant’s recall bias (Shim et al., 2014) can affect data quality. In addition, the FFQ requires participants’ conceptualization skills (Livingstone, Robson, & Wallace, 2004). Thus, both can limit accuracy. Moreover, evidence showed dietary habits varying as a function of a participant’s cultural, ethical, and social background. Therefore, the FFQ must always be tailored to the specific study population (Wakai, 2009).

All in all the FFQ is a self-reported dietary assessment tool capable of assessing how much and what people ate. Table 1 lists the FFQ’s various characteristics. However, studies testing its psychometric quality in terms of reliability and validity show mixed results. Moreover, although, it is possible to use the FFQ under everyday life conditions, considering today’s lifestyle, the traditional paper-pencil version is less convenient for the participant.

1.1.2 24-hour recall

Besides the FFQ, the 24-HR is another self-reported or subjective dietary assessment method (Shim et al., 2014), establishing the most recommended measure for certain populations (Andersen et al., 2011).

For the 24-HR, a participant is asked to recall and describe the previous day’s food intake (Hongu et al., 2015; Kerr, Wright, Dhaliwal, & Boushey, 2015; Shim et al., 2014). As noted by Ngo et al. (2009) the 24-HR can either be self- or interviewer-administered. Several formats are
available for the self-administered 24-HR, such as computerized (Börnhorst et al., 2013, Crispim et al., 2014), web-based (Baranowski et al., 2014), image-based (Arab, Estrin, Kim, Burke, & Goldman, 2011), or camera-assisted (Gemming, Doherty, Kelly, Utter, & Ni Mhurchu, 2013; Gemming et al., 2015a) versions. Generally, 24-HR interviews are widespread, for example as implemented by Gemming et al. (2013) or Gemming et al. (2015a). They can either be performed face-to-face (e.g., Hongu et al., 2015) or by telephone (e.g., Zamora-Ros et al., 2011). For instance, a standardized 24-HR interview can be conducted as follows: First, a participant lists all food consumed. Second, using standardized probes, the interviewer explores whether the participant forgot to report any food. Third, time and eating occasion are identified for each item. Fourth, detailed information on the food and its portion size is collected (Kerr et al., 2015). On the one hand, participants themselves can estimate the food portion size, for example in comparison to a reference such as a standard household measure (Gemming et al., 2015a; Shim et al., 2014). On the other hand, dietitians can quantify the food (Kerr et al., 2015). After the portion size is determined, foods are coded using a food database (Shim et al., 2014). Thus, the 24-HR is capable of determining how much and what people ate. Furthermore, administration and data collection require approximately 20–30 minutes for each day (Börnhorst et al., 2013; Shim et al., 2014), and the 24-HR is often repeated, for example on two (Kerr et al., 2015), three (Gemming et al., 2015a) or six (Hongu et al., 2015) occasions.

In general, data from the 24-HR has been used to assess nutritional adequacy through food, nutrients, and eating habits (Burrows et al., 2010). Specifically, the investigation of cancer and nutrition has been determined as a main field of application for the 24-HR in Europe (Slimani et al., 2011; Zamora-Ros et al., 2011). For this purpose, nutrition and health were monitored.

To analyze its retest reliability, a recent study by Kerr et al. (2015) compared the total energy intake determined by a repeated 24-HR which recorded the energy intake of known food.
A total energy overestimation of 11.3% ($SD = 22.5\%$, $p < .05$) was found for the first 24-HR and one of 10.1% ($SD = 20.8\%$, $p < .05$) for the second when compared against known food.

In view of the 24-HR validity, a study by Gemming et al. (2015a) examined estimated energy intake (EI) by comparing 24-HRs taken with and without camera assistance and total energy expenditure (TEE). While EI was assessed using the 24-HR, TEE was measured using the doubly labeled water method (DLW)$^1$. Evidence suggested that without camera assistance men underestimated TEE by 17% ($p = .001$) and woman by 13% ($p = .001$) and with camera-assistance men underestimated TEE by 9% ($p = .02$) and women by 7% ($p = .004$). Overall, camera assistance reduced underreporting.

A further, recently conducted study on the 24-HR’s validity by Börnhorst et al. (2013), investigated the determinants and prevalence of misreporting in children’s proxy-described 24-HRs. In terms of energy intake, the study revealed an underreporting of 8.0%, plausible reporting of 88.6% and over reporting with 3.4%. Another result showed that the probability of underreporting increased with the child’s BMI, $OR = 1.23$, 95% CI [1.10, 1.37] and age, $OR = 1.19$, 95% CI [1.05, 1.83]. Also, underreporting was higher in low- and medium-income groups, $OR = 1.45$, 95% CI [1.13, 1.86].

Considering the 24-HR’s advantages, its ability to gather quantified and detailed information on food intake (Shim et al., 2014) is favorable. For instance, information on a participant’s cooking practices can be determined (Illner et al., 2012). Mostly, the 24-HR is valued as having a low respondent burden (Arab et al., 2011; Shim et al., 2014), which can partly be attributed to the fact it does not require literacy in its interview-based form. Advantages of the computer- and web-based 24-HR versions range from reduced interviewer effort and lower costs

---

$^1$ DLW = doubly labeled water method; participant ingested labeled water ($^{2}$H$_{2}$O, H$_{2}^{18}$O) at the study baseline. Based on urine samples, the daily TEE is calculated based on changes in $^{18}$O and $^{2}$H.
in data processing to an increased flexibility in the time and location of data collection (Illner et al., 2012).

Beyond these benefits, the 24-HR is limited due to its retrospective format, which potentially induces a recall bias. Data obtained using the 24-HR interview can also be affected by the interviewer’s skills. Moreover, the relative high costs for training interviewers and applying the 24-HR over multiple days are less advantageous (Shim et al., 2014). Also, the operation of multiple interviewers, for instance, seven in the study of Kerr et al. (2015), may increase costs. Furthermore, repeated measurement can influence a participant’s diet (Shim et al., 2014). By requiring detailed food information, the 24-HR can be seen as burdensome for the participant (Baranowski et al., 2014). Finally, the open-ended format requires considerable efforts in data analysis (Shim et al., 2014).

Table 1 lists the 24-HR’s characteristics. On the whole, the 24-HR is suited to assessing how much and what people ate. However, given today’s fast pace of life, a repeated 24-HR following the traditional interview-based format does not seem very comfortable for the participant.

1.1.3 Food photographs

Both, FFQ and 24-HR are frequently used dietary assessment tools (Magarey et al., 2011; Shim et al., 2014) that measure how much and what people ate. Like these two methods, FP have long been widely used in dietary assessment (Korkalo, Erkkola, Fidalgo, Nevalainen, & Mutanen, 2013). However, they differ from the FFQ and 24-HR by focusing more intently on how much people ate or portion size estimation (Korkalo et al., 2013).

FP assists a participant in providing adequate food quantification (Nelson, & Haraldsdóttir, 1998) by depicting a single food item (e.g., Brito, Guimarães, & Pereira, 2014) or composite food
(e.g., Tueni, Mounayar, & Birlouez-Aragon, 2012). For instance, after having consumed a pre-
weighed meal, a participant estimated the portion size by choosing one of several photographs
displaying different weight gradations of the meal (Korkalo et al., 2013). So far, FP studies have
used a varying number of photographs ranging from eight (Nelson, Atkinson & Darbyshire, 1996),
to 22 (Laus et al., 2013), 95 (Brito et al., 2014), 334 (Lazarte et al., 2012), 359 (Bernal-
Orozco et al., 2013), 633 (Tueni et al., 2012) or even 894 (Turconi et al., 2005). Often, the food photographs
were arranged within a booklet, referred to as either a food atlas (Bernal-Orozco et al., 2013) or
something similar (Tueni et al., 2012: photographic atlas; Turconi et al., 2005: food photography
atlas; Venter, MacIntyre, & Vorster, 2000: food portion photograph book; Lazarte et al., 2012:
photo atlas). The food was mostly depicted in three weight gradations (small/medium/large)
(Korkalo et al., 2013; Ovaskainen et al., 2008; Tueni et al., 2012; Venter et al., 2000). Furthermore,
the depicted portions could be graded according to equal increments, for example from the 5th–
95th percentile of the distribution of portion weights observed (Nelson et al., 1996). The national
average portion can be used (e.g., Robson & Livingstone, 2000) as basis if it is known.
Traditionally, FP were presented in a paper-based version (Nelson et al., 1996), whereby the digital
format is common today (e.g., Gauthier et al., 2013; Tueni et al., 2012; Turconi et al., 2005).

Regarding their field of application, FP have been used as a tool for assessing food
consumption (Laus et al., 2013). In addition, FP have been applied in combination with other
methods, such as the FFQ (Brito, Guimarães, & Pereira, 2014) or the 24-HR (Lazarte et al., 2012),
to enhance the food record’s accuracy.

Insight into its psychometric quality can be gained from Laus et al.’s (2013) study on FP’s
retest reliability. Participants were asked to select three food images from a scale of 22 food
photographs to describe 1) what they would like to eat, 2) what they consider healthy, and 3) what
they usually eat. The procedure was repeated one month later. Kappa (K) served as
reliability coefficient, where $K = 1$ indicated perfect retest reliability and $K = 0$ a marginal retest reliability (Rosner, 2010). A reliability of $K = 1.0$ was found among men and $K = .94$ for women asked to choose what they would like to eat. For the total sample $K = 1.0$ was observed for the question of which foods they considered healthy and $K = .71$ for the question of what they usually eat.

A further validity study (Bernal-Orozco et al., 2013) showed that the amount of food estimated using FP differed statistically compared to actual food weight. However, error percentages were considerably lower for FP (2.3%) than food estimation using measuring cups (56.9%, $p = .001$) or food models (325%, $p = .001$).

Previously, Nelson et al.’s (1996) classic study on FP revealed a tendency for overestimating small portions and underestimating large portions. Also, older participants overestimated portion size more frequently. Further results showed nutrients estimated by FP to be within $\pm 7\%$ of actual nutrients.

Advantages of FP include portability when arranged in a food atlas and the variety that can be pretend (Korkalo et al., 2013; Nelson et al., 1996).

However, FP are limited due the fact that gender, age, and BMI can all potentially confound portion size estimation (Nelson et al., 1996). Specifically, these factors can systematically influence portion size estimation and thereby create error variance. Another issue that can be limiting is that estimating portion size requires the participant’s perception, conceptualization, and memory (for more details, see Nelson et al., 1996). Moreover, FP are culture specific (e.g., Lombard, Steyn, Burger, Charlton, & Senekal, 2013) and cannot be applied universally.

FP’ features are summarized in Table 1. All in all, FP appear to be a suitable tool for measuring how much people ate. Even though they can be applied in everyday life conditions, it
does not seem convenient for a participant to estimate his or her consumed portion size using a paper-based food atlas.

1.1.4 Self-reported dietary assessment methods: Summary

As presented above, the FFQ and the 24-HR are frequently used self-reported dietary assessment methods (Magarey et al., 2011; Shim et al., 2014). Although they both assess how much and what people ate, FP concentrate solely on how much. Each method has specific characteristics (Table 1): the FFQ emphasizes frequency of consumption, the 24-HR concentrates on the food intake of the last 24 hours, and FP focuses on quantity.

All of these methods have their particular advantages and disadvantages. For example, while the FFQ can be time- and cost-effective (Shim et al., 2014), it is limited due potential recall bias (Shim et al., 2014). Whereas the detailed information collected by the 24-HR is an advantage (Shim et al., 2014), the short duration creates a limitation. Although FP provide valuable support in portion size estimation (Nelson et al., 1996), gender, age, and BMI potentially confound portion size estimation (Nelson et al., 1996).

Reliability and validity studies are available for FFQ, 24-HR, and FP, respectively (see Table 1). The recent psychometric evidence presented above indicates that although results are mixed, these assessment tools improve measurement of how much and what people ate. While it is possible to apply these three methods under everyday life conditions, filling in a paper-based FFQ, repeatedly joining a 24-HR interview, or carrying around a paper-based FP can be difficult to reconcile with today’s fast paced lifestyle.

In addition to these self-reported dietary assessment methods, gathering a visual food record provides another option for assessing a participant’s dietary intake. The different ways of assessing and coding a visual food record are summarized below.
### Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Focus</th>
<th>Advantages</th>
<th>Limitations</th>
<th>Psychometric Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FFQ</strong></td>
<td>Frequency of consumed food acquired by food list</td>
<td>Time- and cost-effective</td>
<td>Limited food choice, Recall bias, Population specific</td>
<td>Retest reliability: Kolodziejczyk et al., (2012) ( r = .05-.88 )</td>
</tr>
<tr>
<td><strong>24-HR</strong></td>
<td>Food intake of previous 24 hours acquired by interview</td>
<td>Detailed information</td>
<td>Short period of time, Recall bias</td>
<td>Retest reliability: Kerr et al., (2015) ( 1. 24\text{-HR}: 11.3% \pm 22.5% \ (p &lt; .05) \text{ energy overestimation} \ 2. 24\text{-HR}: 10.1% \pm 20.8% \ (p &lt; .05) \text{ energy overestimation} )</td>
</tr>
<tr>
<td><strong>FP</strong></td>
<td>Portion size estimation through comparison with photographs depicting food</td>
<td>Improves portion size estimation</td>
<td>Limited number of foods depicted in photographs, Population specific</td>
<td>Retest reliability: Laus et al., (2013) ( K = 1.0, \text{ when asked to select what they consider as healthy} \ K = 1.0, \text{ when asked to choose what they want to eat (men)} \ K = .94 \ (woman) \ K = .71, \text{ when asked what they usually eat} )</td>
</tr>
</tbody>
</table>

**Note.** Table 1 summarizes the focus, advantages and limitations of the FFQ, 24-HR and FP by means of examples. The measures’ psychometric quality (reliability and validity) is shown by recent studies. FFQ = food frequency questionnaire; 24-HR = 24 hour recall; FP = food photographs; DLW = doubly labeled water method; \( K = \text{Kappa} \)
1.2 A visual food record: Assessment and coding methods

The classic food records discussed above each provide different forms of information: While the FFQ establishes a visual food record in a written form, the 24-HR presents an aural food record and the FP relies on a visual food record in a picture-based format.

The present study, however focuses on a visual food record in a digital, picture-based format that assesses a person’s food intake under everyday life conditions while considering how much and what people ate. Such a visual food record can be assessed by different means. For instance, a self-developed device like the so called e-Button (Jia et al., 2014) or a smartphone which creates a further opportunity to attain a visual food record (Martin et al., 2009a; Six, Schap, Kerr, & Boushey, 2011).

1.2.1 A visual food record based on a self-developed device

Jia et al. (2014) developed the so called e-Button as an electronic device for acquiring a digital, picture-based food record. It is reviewed as a promising new methodology (Illner, Lachat, & Slimani, 2014) and described hereafter.

The e-Button is a small, chest-worn camera. Fixed at a suitable angle on a participant’s shirt, it automatically photographs the food being consumed at two-second interval. In the initial validation study (Jia et al, 2014), seven employees recorded their food intake in the lab during their lunch break. They chose their meal from a predefined list of 100 food samples. If the meal included multiple food items, they placed each item on a separate plate and consumed it separately. To measure portion size, the researcher removed the food before consumption and measured it using the seed displacement method\(^2\). The food was then returned to the participant for consumption.

---

\(^2\) Seed displacement method = The food item is wrapped in plastic and put in a standard container, which is filled with seed. The seed around the food item is measured to determine the food item’s volume
After the data collection, in Jia et al.’s study (2014), food portion size was ascertained through both 1) a software and 2) by coders. In terms of the software, an algorithm first evaluated picture quality and removed the blurred ones. The best one was the manually selected from the remaining pictures. Next, this picture was undistorted using a specific algorithm. After all, the virtual shape model was applied for portion size estimation (Chen, Jia, Li, Sun, & Sun, 2012). The most suitable food shape was chosen from a library and adjusted to fit the depicted food as closely as possible. Using the known plate size as the parameter for spatial information, software estimated the food volume based on the volume of the fitted model. Coders also estimated the food portion size from digital pictures presented on the computer screen (Jia et al., 2014).

Until now, the main field of application for the e-Button has been portion size estimation in dietary assessment (Jia et al., 2014). However, Sun et al. (2014) proposed further areas of application for the e-Button, such as tracking sedentary events and physical activity. For this objective, the type and duration of the respective activity must be identified using acquired pictures.

To gain insight into the e-Button’s psychometric quality, Jia et al.’s (2014) study is considered as an example. As part of this validity study, computer software automatically estimated the food volume of the food in the photographs acquired by the e-Button. Furthermore, three raters estimated the volume from the matching pictures. In addition to these estimations, actual food weight was determined using a physical measure. The results indicated a mean relative error between actual food weight and the weight estimated by the computer software with $M = -2.8\%$ ($SD = 20.4$). Although the mean relative error for rater 1 was $M = 620$ ($SD = 80.0$), it was $M = -13.5$ ($SD = 40.2$) for rater 2 and $M = 34.7$ ($SD = 46.8$) for rater 3. The best mean estimation from the raters was $M = -15.5\%$ ($SD = 41.4\%$) and the worst was $M = 78.8\%$ ($SD = 51.2\%$). All in all, the computer-based method demonstrated less bias and better agreement with the actual food compared with the raters’ estimation (Jia et al., 2014).
Moreover, the e-Button has several advantages resulting from automatic photography. First, a participant’s eating behavior is not interrupted. Second, it is convenient to use, as the participant just needs to turn the e-Button on and off. Third, it does not require the participant’s willingness or ability to take pictures (Jia et al., 2014). Due to Illner et al.’s (2014) study, the e-Button is memory independent and thereby reduces removing recall bias. Additionally, the participant requires neither literacy, large cognitive efforts, nor computer skills.

At the same time, the e-Button has its limitations. For example, only foods with known standard volumes or easily to determine volumes can be used. As such, foods with a shape that changes rapidly, such as ice cream, had to be excluded because they could not be measured accurately by the seed displacement method (Jia et al., 2014). Moreover, certain foods that were an ultimate composite (Jia et al., 2014) or too complex (Illner et al., 2014; Jia et al., 2014) to be identified from an image, also had to be excluded. Concerning the estimated food weight, the computer software errors were large for some foods because 1) the food was not completely displayed, or 2) the food shape was too complex to be fitted to an available shape model. In addition, a volume-to-weight conversion had to be available in the database, otherwise the weight could not be calculated (Jia et al., 2014). As Illner et al. (2014) highlighted, costs for designing the e-Button should be noted, especially if it is to be applied to a large sample. In much the same manner, the data assessment and coding can be financially burdensome.

Table 2 summarizes the e-Buttons characteristics. All in all it appears to be an up-to-date measurement that enables an assessment of how much and what people ate. Although food type can be identified from the acquired food images, its focus to date has been on portion size estimation. Furthermore, due to the above listed reasons it is not applicable for all kinds of foods and thereby limited for implementation under real-life conditions.
1.2.2 A smartphone based visual food record

While the e-Button, as wearable camera, establishes a passive approach to capturing food images, the smartphone based visual food record creates an active method that requires the participant to record food intake (Gemming, et al., 2015b). During recent years, the smartphone has become a frequently used (Rusin et al., 2013) and popular (Kawano & Yanai, 2015) tool for food recording. In fact, smartphone based food recording has been identified as one of the most innovative technologies for measuring food intake (Illner et al., 2012).

To date, several tools are available for assessing a smartphone based visual food record including the FoodLog with image-based assistance (Aizawa et al., 2014), the Nutricam Dietary Assessment Method (NuDAM) (Rollo, Ash, Lyons-Wall, & Russell, 2015), the 24-DR food picture app (Hongu et al, 2015), the Remote Food Photography Method (RFPM) (Martin et al., 2009a; Martin, Kaya, & Gunturk, 2009b; Martin et al., 2012) and the Technology Assisted Dietary Assessment (TADA) (Aflague et al., 2015; Boushey et al., 2015; Schap, Zhu, Delp, & Boushey, 2014), which is based on the mobile telephone food record (mpFR) (Six et al., 2010; Six et al., 2011). All of these smartphone based visual food records operate similarly.

Usually participants received a training on how to handle the smartphone for food recording (e.g., Martin et al., 2012; Rollo et al., 2015; Six et al., 2010). After that, participants captured food selection and leftovers. To remind participants to capture their food intake, sometimes prompts were sent (Martin et al., 2009b; Martin et al., 2012). In some studies, participants were instructed to hold the phone at a 45° angle (Hongu et al., 2015; Rollo et al., 2015) or until a green border appeared on the screen (Aflague et al., 2015) to ensure that all food was visible. In the latter case, the optimal angle was calculated automatically. Moreover, participants were frequently asked to include a fiducial marker (object of known size) in the image to help with reconstructing the environment for subsequent coding. For example, a checkerboard square (Aflague et al., 2015;
Boushey et al., 2015) or an ID card with a specific pattern printed on top (Martin et al., 2009b) served as a color and size reference. Finally, participants sent their visual food record to the researchers’ server via cellular network.

Further aids were employed to increase the completeness of the food record. In a study by Rollo et al. (2015), a structured phone call was made to verify ambiguous food images and to probe for forgotten foods. As an alternative method, participant and a dietician reviewed images by using a 24-HR (Schap et al., 2014). In addition, the participant could shift from an image-based to an text-based mode if the specific food was not contained in the food database (Aizawa et al., 2014).

A smartphone based visual food record cannot solely be assessed in a controlled environment (Aflague et al., 2015; Boushey et al., 2015; Martin et al., 2009a; Martin et al., 2012; Schap et al., 2014; Six et al., 2011) but also under real-life conditions (Aizawa et al., 2014; Hongu et al., 2015; Hutchesson, Rollo, Callister, & Collins, 2015; Martin et al., 2009a; Martin et al., 2012; Rollo et al., 2015). Moreover, the smartphone based visual food record has been studied among different age groups, including 3–10 years (Aflague et al., 2015), 11–15 years (Boushey et al., 2015), 11–18 years (Six et al., 2011), 18–24 years (Aizawa et al., 2014) and 18–65 years (Martin et al., 2012). Likewise, the study duration varied, including two (Aflague et al., 2015; Boushey et al., 2015), three (Martin et al., 2009a; Rollo et al., 2015), six (Martin et al., 2012), seven (Hutchesson et al., 2015), and 14 (Aizawa et al., 2014) days.

After its assessment, the smartphone based visual food record must be coded. This involves identification and quantification of food. Each can be accomplished manually or automatically.

In terms of the manual coding process, a dietician labeled and quantified the food shown in the smartphone based image. Food images and other aids were used as a references for portion size estimation. The food types and the portion size estimation were entered into a nutrient analysis software (Rollo et al., 2015), to provide the nutritional values.
Although primarily performed by coders, automatic food identification is also possible, as shown by Schap et al. (2014). Different mechanisms of automatic food volume estimation are available, including a multi-view volume estimation (Xu et al., 2013), circular referents (Jia et al., 2012), and food-specific shape templates (Chae et al., 2011).

In Schap et al.’s (2015) study, food identification and volume estimation were both conducted automatically. Through segmentation, feature extraction, and classification an image analysis was conducted to enable automated food identification. Next, portion size was estimated automatically based on segmentation and shape template modeling. Finally, the labeled and estimated food was indexed by a nutrient database to obtain detailed nutrient information. Food was identified and automatically estimated in much the same way in a study by Martin et al. (2009b), where reference card detection, food region segmentation and classification along with food amount estimation formed the crucial elements. Because a dietician reviewed the results for possible changes in estimated amounts, the method was referred to as semi-automatic.

As example of the assessment and coding of a smartphone based visual food record, the TADA from Boushey et al., (2015) is depicted in Figure 1. First, the user captured his or her food intake. Next, the digital food images were wirelessly sent to the researchers’ server. There, the captured food and beverage was coded automatically. Afterward, the labeled food images were sent back to the user to correct or confirm the labeling. Then, the volume was estimated and indexed by a nutrient database for nutrient analysis. Finally, the images and the related data were stored for further use.
In view of its field of application, the smartphone-based visual food record was used in epidemiological research (Martin et al., 2012) and for monitoring dietary intake on an individual level via self-monitoring (Hutchesson et al., 2015). Moreover, Rusin et al. (2013) showed in a review of methods for recording dietary intake that mobile phone studies were mainly designed for overweight people (32%), for those with diabetes mellitus (42%), for obese people (45%) and for those who want to maintain their health (10%). For example in the study by Rollo et al. (2015) a smartphone-based visual food study was used among people with the chronic disease type 2 diabetes mellitus.

To determine its psychometric quality, in an initial validation study Martin et al. (2009a) compared the EI estimated by the RFPM with the EI of directly weighed foods over three days under real life and laboratory conditions. Using the RFPM, raters estimated the portion size depicted in the food image by comparing it with standard portion food photographs. The EI was
then calculated with the help of an energy and nutrition database. As the food the participants consumed was pre-weighed, its EI could also be directly measured. The study’s results demonstrated that the RFPM produced reliable estimates ($r = .62$, $p < .0001$) under laboratory conditions and in real-life ($r = .68$, $p < .0001$). In two laboratory based validity tests, the RFPM underestimated energy intake (EI) by -4.7% ($p = .46$) and -5.5% ($p = .76$) respectively, compared with -6.6% ($p = .017$) in real life conditions. Moreover, intra-class correlation (ICC) was .99, 95% CI [.99-.99] for food selection, .91, 95% CI [.87-.94] for plate waste, and .88, 95% CI [.81-.91] for EI. For fat intake the ICC was .92, 95% CI [.88-.94], compared with .85, 95% CI [.77-.89] for carbohydrates and .85, 95% CI [.79-.90] for protein.

In another validation study of the RFPM conducted by Martin et al. in 2012, energy was estimated by DLW and compared with the energy recorded of standardized and customized prompts. For the energy estimation, participants drank water labeled with $^{2}$H$_{2}$O and 2g H$^{18}$O at the beginning of the study. Energy was then calculated on the basis of urine samples and changes in $^{18}$O and $^{2}$H. Of further note, participants receiving customized prompts had additional reminders at personalized meal times. The study’s results showed a significant difference between EI by DLW and estimates by standard prompts with $M = -895$ ($SD = 770$ kcal/day, $P < 0.000$). By contrast, with $M = -270$ ($SD = 748$ kcal/day, $P = 0.22$), there was no significant difference between DLW and estimates by customized prompts.

Rollo et al. (2015) tested the NuDAM validity as another image-based mobile phone dietary record. Energy intake was assessed with NuDAM in the way that raters estimated the portion size of the depicted food with the help of visual aids. The EI was calculated using a nutritional software and TEE was determined by DLW (described above). The results showed that NuDAM significantly underestimation TEE at 0.76 ($SD = 0.20$). Additionally, the following correlations between NuDAM and weighed food records were found: $r = .57$ for energy, $r = .63$, $p < 0.05$ for
Introduction

carbohydrate, $r = 0.78$, $p < 0.01$ for protein, $r_s = .85$, $p < 0.01$ for alcohol, and $r = .24$ for fat. In addition, the interrater reliability (IRR) was tested over 3 days. It ranged from .77 to .99 for the NuDAM based food record.

Moreover, a smartphone based visual food record has many advantages. After participating in the initial study people reported they were willing to use the smartphone based visual food record again (Aflague et al., 2015; Rollo et al., 2015) and for a longer period (Aizawa et al., 2014; Boushey et al., 2015; Rollo et al., 2015). For example, in the study by Boushey et al. (2015) 32% of the participants ($N = 41$) indicated they would use the smartphone based visual food record up to 30 days. In another study, about half of the participants ($N = 45$) considered using a smartphone based recall app for food recording outside the study every day (Hongue et al., 2015). One possible explanation is that participants found the smartphone based food record to be easy to apply. As such, Aflague et al. (2015) showed that 89% of the participants ($N = 63$) confirmed the smartphone based food record was easy to use and Hongue et al (2005) reported this for 96% of the participants ($N = 45$). Further, supporting this finding, Aizawa et al. (2014) showed that additional training significantly increased the perceived ease of use. In general, the majority of participants were able to capture an image of the selected food and leftovers that included all foods (Six et al., 2010). Even children in the age group 3–10 years captured usable images and returned the smartphone (Aflague et al., 2015). Finally, a review by Long et al., (2010) evaluated the smartphone as a tool for reducing burden in dietary assessment. A near real time (vs. recall) food record is another advantage for a smartphone based visual food record.

Some limitations are associated with the smartphone based visual food record. A participant can forget to capture the food intake, causing a misreporting (Martin et al., 2014; Rollo et al., 2015). Moreover, a relationship between time of day and selective reporting was found. In a study by Boushey et al. (2015), participants were more willing to capture their breakfasts and lunches (90%,
respectively) compared with their afternoon (54%) or evening (40%) snacks. Losing the phone or technical problems are further potential factors causing incomplete data (Martin et al., 2014). Handling a smartphone requires specific knowledge which can cause difficulties for elderly people or other less-experienced groups (Long et al., 2010). Moreover, smartphone studies have thus far involved a limited number of participants (Sharp & Allman-Farinelli, 2014), for example the study by Six et al (2011) had 15 participants. Also, population’s representativeness is questionable (Sharp & Allman-Farinelli, 2014). Regarding costs, phones as a data-collection devices are more expensive than are paper-and-pencil based methods (Long et al., 2010). A review by Sharp and Allman-Farinelli (2014) suggested that additional costs can also arise from employing dietitians. Finally, the image analysis of digital food images is described as a critical issue (Steele, 2015).

Accompanied by specific advantages and limitations (Table 2), a smartphone based visual food record is a favorable and up-to-date dietary assessment tool. It is applicable under everyday life conditions for assessing how much and what people ate. Moreover, it was convenient for the participant to use. However, its coding procedure appears to be in the developmental stage.

1.2.3 A visual food record: Summary

As presented above, several innovative methods for obtaining a visual food record are available. Table 2 depicts their respective advantages, limitations and psychometric properties.

The e-Button (Jia et al., 2014), as an example of a self-developed device, allows for the simultaneously assessment of food weight, food type, and food nutrients. Also, the RFPM (Martin et al., 2009a; Martin et al., 2009b; Martin et al., 2012), and mpFR (Six et al., 2010; Six et al., 2011) are frequently used smartphone based visual food records that are suitable for studying food weight, food type and food nutrients. Additionally, they allow for a near real-time dietary assessment and hence prevent a recall bias.
### Table 2

**Characteristics of Visual Food Records**

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Limitations</th>
<th>Psychometric Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-developed device</td>
<td>- Unobtrusive</td>
<td>- Certain foods cannot automatically be identified</td>
<td>Jia et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>- Participant’s willingness not required</td>
<td>- Large weight estimation errors for certain foods by software</td>
<td>Validity</td>
</tr>
<tr>
<td></td>
<td>- No recall bias</td>
<td>- High costs for development, data assessment and coding</td>
<td>Relative error between actual food weight and weight estimated by the software was $M = -2.8%$ ($SD = 20.4$).</td>
</tr>
<tr>
<td>Smartphone</td>
<td>- Ease of use</td>
<td>- Selective misreporting due to time of day</td>
<td>Martin (2009a)</td>
</tr>
<tr>
<td></td>
<td>- Participants showed high willingness to use</td>
<td>- Skills needed to handle smartphone</td>
<td>Validity</td>
</tr>
<tr>
<td></td>
<td>- No recall bias</td>
<td>- Costs for device and coding</td>
<td>Relative errors between actual food weight and estimation by the raters:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Rater 1: $M = 62$, ($SD = 80$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Rater 2: $M = -13.5$ ($SD = 40.2$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Rater 3: $M = 34.7$ ($SD = 46.8$)</td>
</tr>
<tr>
<td>Martin (2009a)</td>
<td></td>
<td></td>
<td>Validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EI by directly weight foods and RFPM under</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Laboratory conditions: $r=.62$, $p&lt;.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Real-life conditions: $r=.68$, $p&lt;.0001$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Underestimation EI by RFPM under</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Laboratory conditions: $-4.7%$ ($p = .46$) (First study), $-5.5%$ ($p = .76$) (Second study)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Real-life conditions: $-6.6%$ ($p = .017$)</td>
</tr>
<tr>
<td>Martin et al. (2012)</td>
<td></td>
<td></td>
<td>Validity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Difference between RFPM and DLW when estimating EI for</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Standardized prompts: $895 \pm 770$ kcal/day, $P &lt; 0.0001$, sig.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Customized prompts: $-270 \pm 748$ kcal/day, $P = 0.22$, n. sig.</td>
</tr>
</tbody>
</table>
### Method

<table>
<thead>
<tr>
<th></th>
<th>Advantages</th>
<th>Limitations</th>
<th>Psychometric Properties</th>
</tr>
</thead>
</table>
| Rollo et al. (2015) | **Validity** | TEE by DLW and NuDAM  
- Sig. underestimation of 0.76 ± 0.20 by NuDAM  
- Correlations between NuDAM and weighted food:  
  - Energy: $r = .57$  
  - Carbohydrates: $r = .63$, $p < 0.05$  
  - Protein: $r = .78$, $p < 0.01$  
  - For alcohol: $r_s = .85$, $p < 0.05$  
  - Fat: $r = .24$  |
|   | **Reliability** | ICC for NuDAM based food record for three raters over three days ICC .77-.99 |

Note. Table 2 summarizes advantages and limitations for the e-Button and smartphone based visual food records by means of examples. The measures psychometric quality are shown by exemplary, recently conducted reliability and validity studies. EI = energy intake; RFPM = remote food photograph method; TEE = total energy expenditure; DLW = doubly labeled water; NuDAM = Nutricam Dietary Assessment Method; ICC = intraclass correlation.
However, both methods possess certain limitations. Even if the e-Button relies on promising technology, its automated food identification and food weight estimation cannot handle with every type of food. Moreover, high costs are related to its development and application (Illner et al., 2014). Consequently, it does not seem applicable to dietary assessment studies using a large population under everyday life conditions. Smartphone based visual food records, however, appear to be easily applicable (Aflague et al., 2015; Hongu et al., 2015) on a daily basis over a relatively long time period, such as for 2 weeks (Aizawa et al., 2014).

Still, the coding manifests as a major issue. Psychometric data for some smartphone based food record are available, but studies on the reliability and validity testing of coding systems for a smartphone based visual food record remain rare. Thus, it is crucial to evaluate existing coding systems for a smartphone based visual food record to ensure reliable measurement.

1.3 Present study

In light of current research and the smartphone based visual food record (Renner et al., 2015) the research questions of the present study are derived.

1.3.1 Summary of the current research

Several fields of research rely on knowledge of a person’s everyday food consumption: How much of what food a person ate and what nutrients that food contained have been revealed as crucial. Today a variety of dietary assessment methods are available. However, none of the available tools focuses on 1) food weight, 2) food identification, and 3) food nutrients at the same time thereby allowing user-friendly data collection under real life conditions while ensuring a high level of psychometric quality.
Common self-reported methods to date include the FFQ, the 24-HR and FP. Although the FFQ and the 24-HR assess both food weight and food type, the FFQ focuses on frequency of consumption (Rusin et al., 2013) and the 24-HR on what was consumed the last 24 hours (Hongu et al., 2015). FP, instead, focus on food volume estimation (e.g., Nelson et al., 1996). Although, each of these methods can be applied under everyday life conditions, they are less convenient in their traditional form for daily use.

Alternatively, options include a visual food record in terms of a self-developed device, such as the e-Button (Jia et al., 2014) or smartphone based food records like the RFPM (Martin et al., 2009a; Martin et al., 2009b; Martin et al., 2012) or the mpFR (Six et al., 2010; Six et al., 2011).

These seem promising because they allow for the simultaneous data collection on food weight, food type, and food nutrients simultaneously. However, since its application is limited to certain foods (Illner et al., 2014), it remains unclear whether the e-Button is viable in a real-life setting. The smartphone based visual food proved easy to apply under everyday life conditions (Hongu et al., 2015) and therefore more appropriate. Yet, this method suffers from a different issue: the coding.

The subjects for coding the visual food record include 1) food weight estimation and 2) food identification. Nutrients are subsequently calculated with a nutritional software that relies on information about the food weight and type. Both food weight estimation and food identification, can be conducted either manually or automatically (e.g., Rollo et al., 2015; Schap et al., 2015). For manual coding, coders label and estimate the food using visual aids such as FP. After coding, food labels and portion sizes are entered into a nutritional software that calculates the nutrients contained (Rollo et al., 2015). However, human ability to estimate portion size accurately limits this approach: it constitutes a problem (Lee et al., 2012) and restricts the smartphone based visual food
record’s efficiency (Martin et al., 2009b). As suggested in Steele’s (2015) review, automated food image analysis can enhance dietary assessment accuracy without the need of relying on coders.

While automated food identification (e.g., Schap et al., 2014) is relatively rare, several mechanisms are available for an automated food volume estimation that relies on multi-view volume estimation, circular referents, or food-specific shape templates (Chae et al., 2011; Jia et al., 2012; 2015; Xu et al., 2013). However, as reviewed by Steele (2015), studies on automated food image analysis showed that the automated systems are not yet sufficiently accurate for dietary assessment. For example, the different ingredients in similar foods cannot be distinguished.

### 1.3.2 Present food record and coding system

Based on recent finding in dietary assessment, Renner et al. (2015) developed a smartphone based visual food record. Participants used a smartphone to record their daily food intake. Their food images were wirelessly sent to the researchers’ server. The participant’s food images were manually coded using a coding system including 1) a food atlas and 2) the nutritional software OptiDiet©. First a coder assigned labels to the depicted food. Next, the coder estimated the food weight by using a digital food atlas. Then, the coder entered the food label and the food weight into OptiDiet©, which calculated the food nutrients.

To date and to the best of my knowledge, this is the first smartphone based visual food record for Germany. Because it simultaneously addresses food weight, food type, and food nutrients while relying on a smartphone based data collection, which is easy to apply under real-life conditions, it seems as a promising approach for dietary assessment. As part of the BMBF-funded research project Eatmotive, SmartFood, a first study using the newly developed smartphone based visual food record was conducted. However, as the reliability of the coding system has yet to be tested, coding accuracy needs to be determined.
1.3.3 Research questions

To address this research gap, the present study evaluated Renner et al.’s (2015) coding system. Specifically, the coding system’s reliability was determined in terms of intercoder reliability. Therefore, two coders (referred to as coder A and B) independently coded all participants’ smartphone based visual food records concerning both food weight and food labeling. Afterward, contained food nutrients were calculated, respectively. Krippendorff’s Alpha (Kalpha) was chosen for the reliability coefficient. By comparing codings it calculated coders’ agreement and thereby estimated intercoder reliability. This study primarily aims to extend knowledge on the following three questions:

- **Research question 1:** To what extent does the present coding systems allow for a reliable estimation of *how much* people ate? To address this question, Kalpha was calculated for all participants and per participant for the estimated weight per meal by coders A and B.

- **Research question 2:** To what amount does the present coding systems allow for a reliable labeling of *what* people ate? Here, Kalpha was calculated for all participants and per participant for the number of assigned labels per meal by coders A and B. Furthermore, the percentage of meals assigned per food category by coders A and B was calculated.

- **Research question 3:** In what way does the present coding systems allow for a reliable determination of the nutrients contained in people’s food? To this end, Kalpha was calculated for different micro- and macronutrients.

To summarize the principal questions: Can the coding system of the newly developed smartphone based visual food record reliably determine *how much* and *what people* ate under real-life conditions?
2. Methods

The study’s participants and the procedure conducted, which relies on the coding system and the data analysis performed, are described in the following.

2.1 Participants

Data were obtained from 10 adults (six males, four females) with a mean age of 51.10 years ($SD = 15.81$, range: 30-74). On average, the participants were overweight (BMI $M = 26.4$, range: 19.55-32.41) and showed 16.44 years of formal education ($SD = 1.66$, range: 13-18). With the exception of one participant who did not specify his years of education, all participants provided complete information on the demographic questions.

The present reliability study’s sample was drawn from the SmartFood study ($N = 99$ participants) conducted by Renner et al., (2015). No significant difference concerning participants’ characteristics were observed between the present study’ and the total SmartFood sample (sex: $\chi^2 (1) = 0.76, p = .38$; age: $t(96) = -0.74, p = .46$; education: $t(93) = -0.95, p = .35$, BMI: $t(96) = -0.66, p = .51$).

The University of Konstanz’s Ethical Review Board approved the SmartFood study, and all participants provided informed consent. Personal information was treated confidentially, and the participants’ anonymity was maintained. As compensation, participants received personalized study feedback.

2.2 Procedure

In the SmartFood study (Renner et al., 2015), participants recorded their entire food intake under everyday life conditions over eight consecutive days. A camera enabled HTC© 6, 7.3 megapixel smartphone served as the electronic device. Participants were asked to photograph their
food at a 45° angle, including a one euro coin as fiducial marker. For each meal, participants were invited to write a food description in an open-question format in their smartphone. If food photographing was not possible (e.g., while driving a car) or inconvenient participants were asked to complete the open-question format as soon as they were able to do so. The digital food photographs and the written food descriptions were wirelessly sent to the researcher’s department server.

As part of the present reliability study of the coding system, two psychologists (subsequently referred to as “coder A” and “coder B”) independently coded the visual food record in terms of 1) food weight and 2) food labels. Coders A and B’s codings were then compared against each other to determine agreement. Finally, the amount of agreement was used to estimate intercoder reliability for an evaluation of the coding system.

2.3 Coding system

A semi-automatic coding system (Renner et al., 2015) was used to analyze the smartphone based visual food record. It consisted of 1) a food atlas and 2) the nutritional software OptiDiet© (version 5, GEO, Linden, Germany). The coding procedure and each component of the coding system are described below.

2.3.1 Coding procedure

After data were collected, a coder manually analyzed each participant’s visual food record using the food atlas and OptiDiet© (Figure 2).
First, the coder chose an appropriate label for the meal shown in the participant’s food record. A depicted meal could either consist of one food item (e.g., tomato soup) or of several food items (e.g., spaghetti Bolognese, parmesan, and salad). To sufficiently describe the meal the coder assigned as many labels as necessary. By entering the label into OptiDiet©, a list of similar labels popped up from which the coder selected the most appropriate. Additionally, the coder estimated the meal’s weight by comparing it with the respective food of known weight displayed in the digital food atlas. If the participant did not photograph the consumed food, the coder used the participant’s written food description provided in the open-question format to choose food labels and determine food weight. Once the food weight and one or more food label(s) had been entered OptiDiet© provided data for over 157 micro- and macronutrients (Renner et al., 2015). In the present reliability study, two coders independently applied this analysis process to all participants’ meals.
2.3.2 Food atlas

The food atlas served as visual aid for estimating the food weight of the participant’s consumed meal. Being a record of habitual food intake complied by department members for the purpose of developing a food atlas (Renner et al., 2015), the atlas is comprised of 41 digital images of 30 types of food.

Seventeen images depicted composite food (e.g., a mixed salad) and 24 single food items (e.g., a banana). For six out of 30 food items, more than one image was available due to a weight gradation. For example, the food item potato was depicted in three weight classes: small, medium, and large (Figure 3). Moreover, the food atlas was divided into six categories: 1) cereals, 2) bread and rolls, 3) vegetables and salads, 4) soups, 5) main dishes, and 6) desserts, cakes, and fruits. For each food image, weight information in grams and the German Federal Food System (GFFS) code was listed (the GFFS is described below) (Hartmann, Bell, Vásquez-Caicedo, Götz, & Brombach, 2006).

![Figure 3. Example of weight gradation in the food atlas. A potato depicted in three weight classes (Renner et al., 2015). Copyright by Renner et al. (2015).](image)

2.3.3 OptiDiet®

To analyze participant’s food intake, the nutritional software OptiDiet® was used for food labeling and nutrient classification. The GFFS served as the underlying food database, providing
nutritional information for approximately 15000 kinds of food. Divided into 19 food categories, the GFFS produced a 9-digit food code for each food item (Hartmann et al., 2006). Each digit offers information about the respective food (Hartmann et al., 2006). For the purpose of this reliability study, only the first digit describing the overall food category was relevant.

2.4 Data analysis

In the present study intercoder reliability was estimated to evaluate the coding system. For this purpose, the amount of agreement between coders A and B’s codings was determined for 1) food weight, 2) food labels (number and content), and 3) food nutrients.

2.4.1 Krippendorff’s alpha

Krippendorff’s alpha (Kalpha) was chosen as the index for intercoder reliability. It can be used to estimate agreement between coders independently coding the same units of analysis (Hayes und Krippendorff, 2007). Kalpha’s underlying mathematical assumption (Krippendorff, 2004) takes the following form:

\[
\text{Agreement} = 1 - \frac{\text{Observed disagreement}}{\text{Expected disagreement}}
\]

By weighing the observed and expected disagreement, Kalpha corrects for disagreement by chance. It ranges from 0 to 1, with 0 indicating no agreement and 1 representing absolute agreement between the codings. Moreover, Kalpha embraces several coefficients and can deal with any number of coders, all metrics, missing data, and a small sample size. Additionally, Kalpha first draws a pairwise (vs. list wise) comparison between the single codings and sums it up afterward (Hayes & Krippendorff, 2007). Kalpha was the best method available as the study’s underlying data were ratio scaled and a pairwise comparison between two codings was needed for content
Methods

All calculations were performed with SPSS (IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.). Although, SPSS does not include a command for calculating Kalpha, a compatible macro was available (Hayes & Krippendorff, 2007). Based on the data characteristics, the macro could be adjusted to account for number of judges, metrics, details, and bootstrapping. In particular, the macro was set as follows: “KALPHA judges CoderA CoderB/level = 4, detail = 0, boot = 0”. The abbreviations “CoderA” and “CoderB” represented coders A and B, “level = 4” stood for a ratio scaled level, and “detail = 0”, and “boot = 0” referred to the bootstrapping method for estimating the distribution’s confidence level.

Even though a bootstrapping of 10000 was recommended (Hayes & Krippendorff, 2007), the macro did not run with this setting. To find a solution for this issue Klaus Krippendorff, the KALPHA index’s inventor, and Andrew Hayes, who wrote the related macro, were contacted. After having discussed the problem, the bootstrapping was determined as insufficient due to the large data range. To make the macro run, the boot was set to zero. Consequently, it was not possible to compute confidence levels for Kalpha with the macro. However, the confidence levels for the mean Kalphas (average of all respective single Kalphas) could be calculated with SPSS at a 95% confidence level.

2.4.2 Excluded cases

Overall, the 10 participants delivered 276 visual food records of their consumed meals. Seven food images were excluded from the entire data analysis for various reasons: No image was provided and the written description said it was a beverage (1×), neither image or description were provided (3×), the image was unclear and no description was provided (1×), one coder partially coded the meal (1×) and one coder did not code the meal (1×).
In total, coder A recognized 920 single food items in all participants’ meals, while coder B identified 866. For adequate content analysis, only the food items labeled with the GFFS code were included, and those labeled with an OptiDiet© specific code were excluded. Subsequently, 5.00% (46/920) of the food items were excluded for coder A and 1.15% (10/866) for coder B. Additionally, one food item was excluded for coder A due to an incomplete GFFS code. All in all, in the content analysis 873 codings were considered for coder A and 856 for coder B.

2.4.3 Overview

To evaluate the coding system, this study tested intercoder reliability by means of Krippendorff’s Alpha. Specifically, agreement between coder A and B’s food weight estimation, number, and content of assigned labels was calculated. Moreover, their agreement on food nutrients was determined. Intercoder reliability was calculated for all participants and per participant, respectively, to establish the coding system’s reliability regarding food weight and food labels. In case of the nutrients Kalpha was calculated for all nutrients and per nutrient.

3. Results

The results on the coding system’s reliability for food weight, food labeling and food nutrients are presented in detail in the following.

3.1 How much people ate: Food weight estimation

Over the course of the study, the 10 participants consumed 269 meals in total. Using the coding system including a food atlas as visual aid (Renner et al., 2015), coders A and B both estimated each meal’s weight.
3.1.1 Food weight estimation for all participants

Coder A and B both estimated the food weight of all participants’ meals (Figure 4).

Figure 4 shows a high agreement for the coders’ food weight estimation for most of the meals and a low agreement for a few meals. Inferential statistics revealed a mean Kalpha of .71, $SD = 0.17$, 95% CI [.59, .84] for the intercoder reliability of food weight estimation for all participants.

Coder A estimated the weight of all 269 meals with $M = 326.68$ g ($SD = 226.12$, $Mdn = 300$ g) and coder B with $M = 297.98$ g ($SD = 202.82$, $Mdn = 270$ g). A $t$-test for independent samples showed no significant difference in their average food weight estimation, $t(536) = 1.55$, $p = .12$, 95% CI [-7.68, 65.08].

Furthermore, for each meal the difference in estimated weight between the coders was calculated over all 269 meals. Figure 5 displays the percentage of meals per difference in estimated weight between the coders for all participants. Same weight estimation and four percentiles are depicted.
Results

For 16 meals (5.95%), both coders estimated food weight identically. For 51 meals (19.96%), the difference in estimated weight between the coders was larger than 0 g but less than or equal to 20.25 g (\(M = 12.57, SD = 5.96\)). For 68 meals (25.28%), the difference was larger than 20.25 g but smaller than or equal to 55 g (\(M = 38.03, SD = 10.09\)). For a further 67 meals (24.91%), the difference was larger than 55 g but less than 118.5 g (\(M = 82.47, SD = 16.57\)). Finally, for 67 meals (24.91%), the difference was larger than 118.5 g but less than or equal to 914 g (\(M = 237.09, SD = 149.98\)). The mean difference in estimated weight between coder A and B was 91.55 g.

\[ \text{Mean difference} = 91.55 \text{ g} \]

3.1.2 Food weight estimation per participant

Moreover, the agreement between the coders’ food weight estimations was calculated per participant (Figure 6).

All in all, the lines representing the coder’s food weight estimation are closer for some participants (e.g., for participant 7) and more distant for others (e.g., for participant 4). This indicates a higher intercoder reliability for some participants and a lower for others. Kalpha ranging from .38 to .88 was found for intercoder reliability for food weight estimation per participant.
Additionally, coder A and B’s total food weight estimation per participant was determined and tested on significance with a $t$-test for independent samples (Table 3). Considering the total food weight estimation, coders A and B both estimated participant 4’s meals as lowest with 5135 g ($M = 233.41$ g, $SD = 114.42$, $Mdn = 247.50$ g) and 4650.8 g ($M = 211.40$ g, $SD = 106.69$, $Mdn = 182$ g), respectively. Also, they estimated participant 2’s meals as highest with 12355 g ($M = 441.25$ g, $SD = 344.63$, $Mdn = 355$ g) and 11191 g ($M = 399.68$ g, $SD = 246.17$, $Mdn = 330$ g). The difference between the total food weight estimation between the coders was smallest for participant 9 with 162.5 g and largest for participant 8 with 2655.7 g. However, the $t$-test did not reveal a significant difference for the means of the total food weight estimation for participant 8 or any other participant.

The difference between the coders’ food weight estimation per participant was also calculated. Figure 7 displays the percentage of meals per difference in estimated weight between coders A and B per participant. Same food weight estimation and four percentiles are shown.

The percentage of meals for which the coders’ food weight estimation was identical was calculated per participant alongside the mean difference in food weight estimation. For more details, see Table 4 (Appendix)

For participants 1, 4, 5, 6, 7, 8 and 10 coders A and B identically estimated 9.68%, 4.54%, 10.34%, 4.17%, 9.52%, 5.71% and 11.11% of the meals’ food weight, respectively. For participants 2, 3 and 9 the coders did not identically estimate the weight of any meal. While the minimum difference in food weight estimation between the coders was 0g, the maximum difference was 914 g (participant 8). The lowest mean of the differences in food weight estimation was found for participant 10 with 58.74 g and the highest for participant 2 with 168.61 g.
Results

Participant 1, $\alpha = .80$

Participant 2, $\alpha = .64$

Participant 3, $\alpha = .49$

Participant 4, $\alpha = .38$

Participant 5, $\alpha = .87$

Participant 6, $\alpha = .79$

Participant 7, $\alpha = .88$

Participant 8, $\alpha = .60$

Participant 9, $\alpha = .80$

Participant 10, $\alpha = .86$

Figure 6. Estimated weight (grams) per meal by coders A and B per participant. Below each graph, Kalpha for the intercoder reliability in estimated weight over all meals per participant is shown. Converging lines indicate a high agreement in the coders’ food weight estimation, while diverging lines indicate a low agreement.
Figure 7. Percentage of meals per difference in estimated weight (grams) between coders A and B per participant (N = 10). Higher bars indicate a larger amount of meals for the respective weight difference, while lower bars indicate a smaller amount.
### Table 3

**T-Test for the Food Weight Estimation**

<table>
<thead>
<tr>
<th>Partic</th>
<th>Coder A</th>
<th></th>
<th></th>
<th>Coder B</th>
<th></th>
<th></th>
<th></th>
<th>t-test</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>M</td>
<td>SD</td>
<td>Mdn</td>
<td>Total</td>
<td>M</td>
<td>SD</td>
<td>Mdn</td>
<td>Dif.</td>
<td>N</td>
<td>df</td>
<td>t</td>
<td>p</td>
<td>95% CI</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10355.8</td>
<td>334.06</td>
<td>251.33</td>
<td>295</td>
<td>9892.6</td>
<td>319.12</td>
<td>222.71</td>
<td>264</td>
<td>463.2</td>
<td>31</td>
<td>60</td>
<td>.25</td>
<td>.81</td>
<td>[-105.70, 135.59]</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>12355</td>
<td>441.25</td>
<td>344.63</td>
<td>355</td>
<td>11191</td>
<td>399.68</td>
<td>246.17</td>
<td>330</td>
<td>1164</td>
<td>28</td>
<td>54</td>
<td>.52</td>
<td>.61</td>
<td>[-118.9, 202.04]</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9405</td>
<td>408.91</td>
<td>230.97</td>
<td>380</td>
<td>8375</td>
<td>364.13</td>
<td>194.83</td>
<td>350</td>
<td>1030</td>
<td>23</td>
<td>44</td>
<td>.71</td>
<td>.48</td>
<td>[-82.20, 171.77]</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5135</td>
<td>233.41</td>
<td>114.42</td>
<td>247.50</td>
<td>4650.8</td>
<td>211.40</td>
<td>106.69</td>
<td>182</td>
<td>484.2</td>
<td>22</td>
<td>42</td>
<td>.66</td>
<td>.51</td>
<td>[-45.30, 89.32]</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8210</td>
<td>283.10</td>
<td>184.69</td>
<td>225</td>
<td>8615</td>
<td>297.07</td>
<td>227.43</td>
<td>205</td>
<td>405</td>
<td>29</td>
<td>56</td>
<td>-2.6</td>
<td>.01</td>
<td>[-122.95, 95.02]</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>11849</td>
<td>493.71</td>
<td>182.16</td>
<td>451</td>
<td>10444.5</td>
<td>435.19</td>
<td>154.70</td>
<td>401.5</td>
<td>1404.5</td>
<td>24</td>
<td>46</td>
<td>1.2</td>
<td>.24</td>
<td>[-39.68, 156.72]</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6387</td>
<td>304.14</td>
<td>181.10</td>
<td>295</td>
<td>5359.5</td>
<td>255.21</td>
<td>141.51</td>
<td>195</td>
<td>1027.5</td>
<td>21</td>
<td>40</td>
<td>.98</td>
<td>.36</td>
<td>[-52.44, 150.29]</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10190</td>
<td>291.14</td>
<td>199.17</td>
<td>300</td>
<td>7534.3</td>
<td>215.27</td>
<td>1555.8</td>
<td>165</td>
<td>2655.7</td>
<td>35</td>
<td>68</td>
<td>1.78</td>
<td>.08</td>
<td>[-9.42, 161.17]</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>6182</td>
<td>309.10</td>
<td>141.71</td>
<td>335</td>
<td>6019.5</td>
<td>300.98</td>
<td>156.00</td>
<td>302</td>
<td>162.5</td>
<td>20</td>
<td>38</td>
<td>.17</td>
<td>.86</td>
<td>[-87.28, 103.53]</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7808.5</td>
<td>216.90</td>
<td>178.29</td>
<td>125</td>
<td>8075</td>
<td>224.31</td>
<td>216.81</td>
<td>137.5</td>
<td>266.5</td>
<td>36</td>
<td>70</td>
<td>-.58</td>
<td>.88</td>
<td>[-100.71, 85.90]</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Table 3 shows the food weight estimation per participant by coders A and B. A *t*-test for independent samples was used to determine the statistical difference in the coders' mean food weight estimations per participa.
3.2 What people ate: Food labeling

Besides analyzing *how much* people ate, the question of *what* people ate was addressed. Both coders labeled each meal using the coding system (Renner et al., 2015) comprising the nutritional software OptiDiet©. Their agreement regarding 1) the number and 2) the content of the assigned labels was calculated.

3.2.1 Number of assigned food labels for all participants

Figure 8 shows the number of labels assigned by coders A and B for all participant’s meals.

![Graph showing number of labels per meal for coders A and B.](image)

*Figure 8.* Number of assigned labels per meal (*N* = 269) by coders A and B for all participants (*N* = 10). While converging lines demonstrate a high agreement in the number of assigned labels between the coders, diverging lines show a low agreement.

Descriptively, Figure 8 displays high agreement between coders A and B for the number of assigned labels for most of the meals and a low agreement for a few meals. For the intercoder reliability of the number of assigned food labels for all participants, the inferential statistics showed a mean Kalpha of .76, *SD* = 0.11, 95% CI [.69, .84].

Coder A assigned *M* = 3.47 (*SD* = 2.15, *Mdn* = 3) labels to each of the 269 meals and coder B assigned *M* = 3.22 (*SD* = 1.86, *Mdn* = 3) labels. A *t*-test for independent samples revealed a difference of .25 labels in their average assignment of labels as not significant, *t* (536) = 1.44, *p* = .15, 95% CI [-.09, .59].
Moreover, the frequency of agreement in the number of assigned labels was analyzed over all meals. Figure 9 displays the percentage of meals per difference in the number of assigned labels between coders A and B for all participants.

For 149 meals (55.4%), coders A and B assigned an identical number of labels. For 75 meals (27.9%), the number of assigned labels differed by 1, for 31 meals (11.5%) by 2 and for eight meals (3%) by 3. For three meals (1.1%), the number of assigned labels differed by 4; for two meals (0.7%) by 5; and for one meal (0.4%) by 6. The mean difference in the number of assigned labels between the coders was .70 labels.

### 3.2.2 Number of assigned food labels per participant

The agreement in the number of assigned labels by coders A and B was also determined per participant (Figure 10).

Overall, the lines presenting the number of assigned labels by coders A and B are closer for some participants (e.g., for participant 4) and more distant for others (e.g., for participant 3). This indicates that intercoder reliability is higher for some participants and a lower for others. Inference
Results

Statistics revealed a Kalpha ranging from .55 to .92 for intercoder reliability in the number of assigned labels per participant.

Moreover, the total number of assigned labels by coders A and B per participant was calculated. A t-test for independent samples was used to test the difference in the coders’ assignment of labels on significance (Table 5).

Examining the total number of assigned labels, the coders assigned the lowest number to participant 4’s meals with 64 labels \((M = 2.91, SD = 1.93, Mdn = 2.0)\) and 58 \((M = 2.64, SD = 15, Mdn = 2.5)\), respectively. Both, coder A and B assigned the highest number to participant 6’s meals with 130 labels \((M = 5.42, SD = 1.79, Mdn = 6.0)\) and 117 labels \((M = 4.88, SD = 1.87, Mdn = 5.0)\). The difference in the number of assigned labels between coders A and B was the largest for participant 3 with 19 labels. Nevertheless, the t-test for independent samples did not show a significant difference for the total number of assigned labels for any participant.

Furthermore, the difference in the number of assigned labels between coders A and B were calculated for each participant. Figure 11 shows the percentage of meals per difference in the number of assigned labels between the coders per participant.

Moreover, the percentage of meals for which the coders assigned the same number of labels was calculated per participant alongside the mean difference in the number of assigned labels (Table 6).

The difference in the number of assigned labels was zero in 34.78% (minimum) for the meals of participant 3’s meals and 71.43% (maximum) for participant 8’s meals. The lowest mean of the differences in the number of assigned labels was found for participant 7 (.33 labels) and the highest for participant 3 (1.26 labels).
Results

Participant 1, $\alpha = .73$

Figure 10. Number of assigned labels per meal ($N = 269$) by coders A and B per participant ($N = 10$). Below each graph, Kalpha for the intercoder reliability in the number of assigned labels over all meals per participant is shown. While converging lines indicate a high agreement in the number of assigned labels, diverging lines indicate a low agreement.
### Table 5

**T-test for the Number of Assigned Labels**

<table>
<thead>
<tr>
<th>Partic</th>
<th>Coder A</th>
<th></th>
<th></th>
<th></th>
<th>Coder B</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>t-test</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total M</td>
<td>SD</td>
<td>Mdn</td>
<td>Total M</td>
<td>SD</td>
<td>Mdn</td>
<td>Diff.</td>
<td>N</td>
<td>df</td>
<td>t</td>
<td>p</td>
</tr>
<tr>
<td>1</td>
<td>82</td>
<td>2.65</td>
<td>1.41</td>
<td>2.0</td>
<td>90</td>
<td>2.90</td>
<td>1.49</td>
<td>3.0</td>
<td>8</td>
<td>31</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>101</td>
<td>3.61</td>
<td>3.06</td>
<td>3.0</td>
<td>84</td>
<td>3.00</td>
<td>1.98</td>
<td>2.5</td>
<td>17</td>
<td>28</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td>98</td>
<td>4.26</td>
<td>2.26</td>
<td>4.0</td>
<td>79</td>
<td>3.43</td>
<td>1.73</td>
<td>3.0</td>
<td>19</td>
<td>23</td>
<td>44</td>
</tr>
<tr>
<td>4</td>
<td>64</td>
<td>2.91</td>
<td>1.93</td>
<td>2.0</td>
<td>58</td>
<td>2.64</td>
<td>1.50</td>
<td>2.5</td>
<td>6</td>
<td>22</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>112</td>
<td>3.86</td>
<td>2.17</td>
<td>4.0</td>
<td>112</td>
<td>3.86</td>
<td>1.55</td>
<td>4.0</td>
<td>0</td>
<td>29</td>
<td>56</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
<td>5.42</td>
<td>1.79</td>
<td>6.0</td>
<td>117</td>
<td>4.88</td>
<td>1.87</td>
<td>5.0</td>
<td>13</td>
<td>24</td>
<td>46</td>
</tr>
<tr>
<td>7</td>
<td>77</td>
<td>3.67</td>
<td>1.56</td>
<td>3.0</td>
<td>70</td>
<td>3.33</td>
<td>1.39</td>
<td>3.0</td>
<td>7</td>
<td>21</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>99</td>
<td>2.83</td>
<td>2.01</td>
<td>2.0</td>
<td>97</td>
<td>2.77</td>
<td>2.03</td>
<td>2.0</td>
<td>2</td>
<td>35</td>
<td>68</td>
</tr>
<tr>
<td>9</td>
<td>76</td>
<td>3.80</td>
<td>1.67</td>
<td>4.0</td>
<td>65</td>
<td>3.25</td>
<td>1.41</td>
<td>3.5</td>
<td>11</td>
<td>20</td>
<td>38</td>
</tr>
<tr>
<td>10</td>
<td>95</td>
<td>2.64</td>
<td>1.76</td>
<td>2.0</td>
<td>95</td>
<td>2.64</td>
<td>2.18</td>
<td>2.0</td>
<td>0</td>
<td>36</td>
<td>70</td>
</tr>
</tbody>
</table>

**Note.** Table 5 shows the number of assigned labels per participant by coders A and B. A t-test for independent samples was used for the difference in the coders’ mean number of assigned labels per participant.
Figure 11. Percentage of meals per difference in the number of assigned labels between coders A and B per participant (N = 10). While higher bars indicate a larger amount of meals for the respective difference in the assigned labels, lower bars indicate a smaller amount.
Table 6

**Differences in the Number of assigned Labels**

<table>
<thead>
<tr>
<th>Partic</th>
<th>No of meals</th>
<th>M difference</th>
<th>Assigned labels</th>
<th>Zero</th>
<th>One</th>
<th>Two</th>
<th>Three</th>
<th>Four</th>
<th>Five</th>
<th>Six</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>31</td>
<td>.65</td>
<td>15</td>
<td>48.39</td>
<td>12</td>
<td>38.71</td>
<td>4</td>
<td>12.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>1.04</td>
<td>14</td>
<td>50.00</td>
<td>6</td>
<td>21.43</td>
<td>5</td>
<td>17.86</td>
<td>1</td>
<td>03.57</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>1.26</td>
<td>8</td>
<td>34.78</td>
<td>8</td>
<td>34.78</td>
<td>3</td>
<td>13.04</td>
<td>2</td>
<td>8.7</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>.45</td>
<td>13</td>
<td>59.10</td>
<td>8</td>
<td>36.36</td>
<td>1</td>
<td>4.55</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
<td>.69</td>
<td>18</td>
<td>62.07</td>
<td>4</td>
<td>13.79</td>
<td>5</td>
<td>17.24</td>
<td>2</td>
<td>6.9</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>.88</td>
<td>11</td>
<td>45.83</td>
<td>8</td>
<td>33.33</td>
<td>4</td>
<td>16.66</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>.33</td>
<td>14</td>
<td>66.66</td>
<td>5</td>
<td>23.81</td>
<td>2</td>
<td>9.52</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>.46</td>
<td>25</td>
<td>71.43</td>
<td>6</td>
<td>17.14</td>
<td>2</td>
<td>5.71</td>
<td>2</td>
<td>5.71</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>.65</td>
<td>11</td>
<td>55.00</td>
<td>6</td>
<td>30</td>
<td>2</td>
<td>10</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>36</td>
<td>.61</td>
<td>20</td>
<td>55.55</td>
<td>12</td>
<td>33.33</td>
<td>3</td>
<td>8.33</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note.* Table 6 presents the mean difference in the number of assigned labels between coders A and B per participant. The exact difference in the number of assigned labels between the coders per participant is also depicted.


### 3.2.3 Content of assigned food labels

As the second part of the food label analysis, the assigned labels’ content was evaluated. For this purpose the GFFS, with its 19 food categories, was adopted for the structure. The frequency each coder assigned a food item to each of the food categories was determined. Figure 12 displays the food categories and the percentage of food items assigned by coders A and B.

![Figure 12](image.png)

*Figure 12. Percentage of food items assigned per food category by coders A and B. Longer bars indicate a larger amount of food items per food category, while shorter bars demonstrate a smaller amount.*
The number of food items the coders assigned to each food category varied. As a minimum, coder A and B assigned 0% and .11% of the food items to category J (Vegetarian food), respectively and 18.10% and 20.21% to category X (Components primarily vegetable) as maximum.

From the total of 269 meals, coder A identified 873 single food items over all meals and coder B 856. Table 7 shows the number and percentage of food labels assigned by coders A and B to each of the 19 GFFS food categories. A $\chi^2$-test was calculated to test if there was a significant difference in the number of assigned labels between the coders in any food category.

Table 7

<table>
<thead>
<tr>
<th>GFFS food category</th>
<th>Total number and percentage of assigned labels</th>
<th>$\chi^2$-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coder A</td>
<td>Coder B</td>
</tr>
<tr>
<td>B (Bread)</td>
<td>85 (9.74%)</td>
<td>81 (9.46%)</td>
</tr>
<tr>
<td>C (Cereals and rice)</td>
<td>30 (3.44%)</td>
<td>46 (5.38%)</td>
</tr>
<tr>
<td>D (Pastry products)</td>
<td>24 (2.75%)</td>
<td>28 (3.27%)</td>
</tr>
<tr>
<td>E (Egg and pasta products)</td>
<td>8 (.92%)</td>
<td>11 (1.29%)</td>
</tr>
<tr>
<td>F (Fruit)</td>
<td>85 (9.74%)</td>
<td>87 (10.16%)</td>
</tr>
<tr>
<td>G (Vegetables)</td>
<td>115 (13.17%)</td>
<td>97 (11.33%)</td>
</tr>
<tr>
<td>H (Legumes, nuts and seeds)</td>
<td>24 (2.75%)</td>
<td>18 (2.11%)</td>
</tr>
<tr>
<td>J (Vegetarian food)</td>
<td>1 (.11%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>K (Potatoes, starch-rich plant shares, mushrooms)</td>
<td>15 (1.72%)</td>
<td>17 (1.99%)</td>
</tr>
<tr>
<td>M (Milk and cheese)</td>
<td>124 (14.20%)</td>
<td>120 (14.02%)</td>
</tr>
<tr>
<td>N (Nonalcoholic drinks)</td>
<td>3 (.34%)</td>
<td>5 (.58%)</td>
</tr>
<tr>
<td>O (Oils, fats, butter, sour cream)</td>
<td>38 (4.35%)</td>
<td>30 (3.50%)</td>
</tr>
<tr>
<td>R (Spices, seasonings, additives)</td>
<td>10 (1.15%)</td>
<td>5 (.58%)</td>
</tr>
<tr>
<td>S (Sweets, sugar, chocolate, ice cream)</td>
<td>43 (4.93%)</td>
<td>36 (4.21%)</td>
</tr>
<tr>
<td>T (Fish and seafood)</td>
<td>8 (.92%)</td>
<td>10 (1.17%)</td>
</tr>
<tr>
<td>U (meat)</td>
<td>12 (1.37%)</td>
<td>8 (.93%)</td>
</tr>
<tr>
<td>W (Sausage and other meat products)</td>
<td>43 (4.93%)</td>
<td>39 (4.56%)</td>
</tr>
<tr>
<td>X (Components primarily vegetable)</td>
<td>158 (18.10%)</td>
<td>173 (20.21%)</td>
</tr>
<tr>
<td>Y (Components primarily animal product)</td>
<td>47 (5.39%)</td>
<td>45 (5.26%)</td>
</tr>
</tbody>
</table>

Note. Table 7 presents the total number and percentage of food items assigned to each of the 19 GFFS food categories by coders A and B. The statistical significance between the coder’s assigned number of labels per category was calculated using a $\chi^2$-test.
The $\chi^2$-test showed none of the difference in the number of assigned labels between the coders in any food category as significant. One exception was category “C (Cereals, rice)”, which was revealed as marginally significant, $\chi^2(1) = 3.86, p = .0049$.

### 3.3 What people’s food consisted of: Food nutrients

In addition to the food weight estimation and the assigned food labels, the food nutrients were analyzed. They were calculated with the nutritional software OptiDiet© as part of the coding system developed by Renner et al. (2015), based on coders A and B’s food weight estimation.

#### 3.3.1 Reliability for all nutrients

To obtain insight into the coders’ agreement on a micro level, Kalpha was calculated for the nutrients contained in the participants’ meals. For all 157 nutrients (OptiDiet© output) and over all meals an intercoder reliability of $\alpha = .67, SD = 0.08, 95\% CI [.59, .72]$ was found.

#### 3.3.2 Reliability per nutrient

Moreover, Kalpha was determined per micro- and macronutrient. Figure 13 shows the intercoder reliability for 1) all nutrients, 2) the group of micro- and macronutrients, and for 3) each micro- and macronutrient. While Figure 13 only displays Kalpha, Table 8 provides additional information, such as how many of the 157 nutrients (OptiDiet© output) belong to each nutrient, respectively.
The intercoder reliability for the micronutrients was $\alpha = .67$, $SD = 0.03$, 95% CI [.58, .72] and the intercoder reliability for macronutrients was $\alpha = .66$, $SD = .10$, 95% CI [.39, .95]. Alcohol showed the lowest reliability within the group of macronutrients with $\alpha = .56$, $SD = 0.01$, 95% CI [.55, .58] and water the highest with $\alpha = .79$, $SD = 0.01$, 95% CI [.79, .79]. In the macronutrient group, vitamins showed the lowest intercoder reliability with $\alpha = .63$, $SD = 0.09$, 95% CI [.59, .67] and organic acids the highest with $\alpha = .68$, $SD = 0.02$, 95% CI [.65, .71].
Table 8

Kalpa For Nutrients

<table>
<thead>
<tr>
<th>Nutrient</th>
<th>Kalpa</th>
<th>N</th>
<th>SD</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minerals</td>
<td>.69</td>
<td>14</td>
<td>0.05</td>
<td>[.66, .72]</td>
</tr>
<tr>
<td>Vitamins</td>
<td>.63</td>
<td>19</td>
<td>0.09</td>
<td>[.59, .67]</td>
</tr>
<tr>
<td>Organic acids</td>
<td>.68</td>
<td>5</td>
<td>0.03</td>
<td>[.65,.71]</td>
</tr>
<tr>
<td>Carbohydrates</td>
<td>.64</td>
<td>38</td>
<td>0.07</td>
<td>[.62, .66]</td>
</tr>
<tr>
<td>Fats</td>
<td>.61</td>
<td>53</td>
<td>0.13</td>
<td>[.58, .65]</td>
</tr>
<tr>
<td>Water</td>
<td>.79</td>
<td>1</td>
<td>0.00</td>
<td>[.79, .79]</td>
</tr>
<tr>
<td>Alcohol</td>
<td>.56</td>
<td>3</td>
<td>0.01</td>
<td>[.55, .58]</td>
</tr>
<tr>
<td>Proteins</td>
<td>.72</td>
<td>24</td>
<td>0.06</td>
<td>[.70, .75]</td>
</tr>
</tbody>
</table>

*Note.* Table 8 presents Kalpha for each nutrient. In addition, it is shown how many of the $N = 157$ nutrients (OptiDiet© output) belong to each nutrient.

4. Discussion

The primary aim of this study was to investigate the reliability of the coding system of Renner et al.’s (2015) newly developed smartphone based visual food record. In Renner et al.’s (2015) SmartFood study, participants used a smartphone to capture their dietary intake. Based on a coding system, including a digital food atlas and the nutritional software OptiDiet©, a coder analyzed the participant’s visual food record (Renner et al., 2015). To determine the coding system’s reliability, in the present study, two coders independently coded all of the participant’s food images. Kalpha was calculated to investigate their agreement for 1) food weight, 2) food labels, and 3) food nutrients.

The main results of this study show the coding system’s overall reliability in establishing food weight, food labels, and food nutrients for all participants. However, the coding system’s reliability varied greatly for single participants or single nutrients. Although reliability was high for some participants or nutrients, it was medium or low for others. Each finding is discussed in detail below.
4.1 The coding system’s reliability for food weight

Since they have different implications, the coding system’s reliability for food weight estimation for all participants and per participant are discussed separately. Additionally, ideas for future research are provided.

4.1.1 Reliability for food weight for all and per participant

An intercoder reliability of $K_{alpha} = .71$ was found for the food weight estimation for all participants’ meals estimated by coders A and B. This means that the two coders agreed on approximately 70% of their overall food weight estimation. To further interpret this reliability outcome, Krippendorff’s (2013) recommendations for classifying $K_{alpha}$ were considered. While one can rely on variables with a $K_{alpha}$ higher than .80, conclusions from variables with a $K_{alpha}$ between .67 and .80 should be drawn with caution (Krippendorff, 2013). In accordance with these guidelines, one may conclude, that a $K_{alpha}$ under .67 can be classified as low, a $K_{alpha}$ larger than .67 but smaller than or equal to .80 as medium, and a $K_{alpha}$ above .80 as high. In this light, the coding system’s reliability for food weight estimation can be classified as medium.

Furthermore, Krippendorff (2013) recommended setting the reliability standard in relation to the research question. The research question of the present study investigates the coding system’s reliability for food weight estimation. Previous studies described food weight or portion size estimation as challenging (Lillegaard, Øverby, & Andersen, 2005) and as providing one of the major errors in measuring food intake (Nelson et al., 1996). Since food weight estimation requires several brain functions, such as perception, conceptualization, and memory (Nelson et al., 1996), and because these are presumably as prone to error as any other human function, it seems questionable whether a reliability of 1.0 is a realistic optimum when determining $K_{alpha}$ for food weight estimations. In line with Krippendorff’s (2013) guidelines for interpreting $K_{alpha}$, the
coding system’s reliability for food weight estimation is medium to high. However, taking the limitations of the nature of manual food weight estimation into account, the reliability found seems high.

Since the coding system mainly relies on the food atlas as a visual aid for estimating food weight, the found reliability for the coding system’s food weight estimation can be attributed to the food atlas. Hence, the food atlas can be seen as highly reliable component of the coding system for food weight estimation. This finding is consistent with previous research, which shows that food photographs improve food portion size estimation (e.g., Bernal-Orozco et al., 2013; Lombard et al., 2013; Ovaskainen et al., 2008).

Also, no significant difference was found between coder A’s and coder B’s total food weight estimation. The coders attributed the exact same weight to a certain percentage of the meals and only a small difference to a high percentage of meals. Both findings indicate a high agreement between the coders’ estimations. This can, in turn, be attributed to the coding system, allowing for a highly reliable food weight estimation.

Considering the coding system’s reliability for food weight estimation the intercoder reliability for the ten participants ranged from $K_{alpha} = .38$ to .88. That means that intercoder agreement when estimating food weight was 38% for one participant (minimum), 88% for another participant (maximum), and between these two extremes for the other eight participants. Taking Krippendorff’s (2013) recommendations into account, the coding system’s reliability for food weight estimation was high for some participants but medium and low for others.

Regarding the total food weight estimation, coders A and B both estimated the food weight for the same two participants as the highest and lowest, respectively. Also, no significant difference in the total food weight estimation between the coders was found for any participant. Other results
revealed that the coders estimated the food weight for the meals of some participants identically, but not for other participants.

Because the coding system’s overall reliability can be evaluated as high, it seems qualified for studies investigating food weight on a global level. For instance, the system could be applied to compare how much people eat between different populations, such as a clinical sample with an eating disorder and a non-clinical sample without an eating disorder.

All in all, the evidence for the coding system’s reliability of food weight estimation per participant was mixed. Hence, the coding system does not seem ready yet to be applied in studies focusing on how much people ate on an individual level. For instance, it cannot currently be recommended to employ the coding system in a study which aims to provide individual feedback on the food weight consumed by participants with an eating disorder.

4.1.2 Future research on the coding system’s reliability for food weight

Although, the food atlas appeared as a highly reliable component of the coding system for food weight estimation, the food atlas suffers from some limitations. It relies on a relatively small number of food images, namely 41. On the contrary, for instance, Turconi et al.’s (2005) food atlas included approximately 900 images. The fact that the food atlas’s photographs were not completely standardized creates another issue. With the objective of improving the coding system’s reliability for food weight estimation, future research on extending the food atlas is needed. In the present study a first study was conducted to extend and standardize the food atlas. To extend the food atlas developed by Renner et al. (2015), additional food images were captured in the present study. To ensure high image quality, the food photographed in approximately half of the images was not real, but fake (plastic). Similar to previous research (e.g., Turconi et al., 2005) a laboratory room was set up as a photo studio including a light, a reflector, and a digital SLR
camera. As in past studies (e.g., Korkalo et al., 2013; Robson & Livingstone, 2000; Turconi et al., 2005), the camera was mounted on a tripod and set up at a 45° angle to optimally capture the food’s height and depth. Also, in accordance with previous studies (e.g., Vereecken, Dohogne, Covents, & Maes, 2010), the dishes were standardized (white plate, small glass bowl, large glass bowl) and placed on a white surface with a white background. As fiducial marker, a one Euro coin was included in every picture. To achieve highly standardized pictures, the position of all equipment, including the dishes were fixed during the procedure of food photographing.

To standardize the kind of food items depicted in the food atlas, the GFFS’s (Hartmann et al., 2006) 19 food categories were used. While the original food atlas (Renner et al., 2015) included six food categories, the one extended in the present study comprised of 19. To establish each food category, an appropriate number of food items were chosen on the empirical basis of Renner et al.’s (2015) Smartfood study. More precisely, the food items representing each food category were chosen in accordance with their frequency of consumption in the Smartfood study. Moreover, composite meals were no longer depicted (with the exception of 11 meals), since the coders perceived the labeling of composite foods as creating error variance. Therefore, the extended food atlas was primarily based on photographs of single food items. Furthermore, each food item was displayed in three weight gradations (small/medium/large). To define a small, medium, or large portion of a specific food, first, the food’s standard portion size was determined according to the recommendations of the Deutsche Gesellschaft für Ernährung (DGE, 2015). The standard portion size was set as the medium portion size. Following the recommendations of a dietician, the small and large portion size were obtained by subtracting or adding 30% of the medium portion size, respectively. However, as the small and large portion sizes did not visually differ from the medium for several food items, an individualized portion size gradation was photographed for each of these
items. Overall, the original food atlas (Renner et al., 2015) was extended from 30 to 117 with the total number of images increasing from 41 to 417 images.

For a visual impression of the standardization see Figure 14. The image on the left (14a) shows the meal “Spaghetti Bolognese” in the original atlas (Renner et al., 2015), and the image on the right (14b) depicts the same meal in the extended food atlas. As mentioned above, the surface, background, dishes, and photo angle were standardized in the extended version to increase accuracy.

Although, the original food atlas (Renner et al., 2015) was extended in this study, future research is required to further enlarge the food atlas with more food images. Moreover, it is necessary to test whether the reliability of the food weight estimation with the coding system increases with the increased number of food images depicted in the food atlas.

Apart from enhancing the food atlas further, future research should investigate other causes of limited reliability of the food weight estimation of some participants on an individual level. For this purpose, it would be worthwhile to manually analyze all of the food images, for the participants for whom the coding system showed limited food weight reliability. Analysis of their food images could provide explanations for the low reliability.
From the coding experience it can be theoretically inferred that the following conditions can conceivably limit accuracy of food weight estimation using visual food records: 1) The single food components were difficult or impossible to define, or 2) the food components were difficult to estimate and one coder was better able to cope with this. Specially, the single food components of a composite meal can be difficult to recognize (e.g., in a mixed salad), sometimes because they are not clearly visible (e.g., cereal in milk). Other potential factors that can reduce a single food item’s recognizability are blurry images or images taken from an angle that does not adequately show the food’s height and depth. Moreover, some foods’ weight appears to be more challenging to estimate than others, for instance, margarine (e.g., Nelson et al., 1996). Finally, the participants for whom the coding system showed limited reliability in food weight estimation may have consumed meals that were not displayed in the food atlas therefore difficult to estimate.

4.2 The coding system’s reliability for food labels

As for the food weight, the coding system’s reliability is discussed for the number of assigned labels for all participants and per participant. Additionally, the content of the assigned food labels was analyzed on the overall level. Finally, recommendations for future research are provided.

4.2.1 Reliability for food labels for all and per participant

For the number of assigned food labels for all participants an intercoder reliability of $K_{alpha} = .76$ was observed. This coefficient indicates that the coders agreed in approximately 76% of overall number of labels assigned. Considering Krippendorff’s (2013) guidelines, this reliability can be classified as medium to high. In light of the research question coders’ agreement of approximately 80% in their assignment of food labels to 269 meals can be evaluated as excellent.
Thus, the coding system appears to be a tool that enables highly reliable coding in terms of the number of assigned labels.

Another result showed that both coders assigned an identical number of labels for approximately 50% of all participants’ meals. In about 30% of the meals, their number of assigned labels varied only by 1 label and in about 10% the number differed by 2 labels. The maximum number of labels the coders’ assignments differed by was 6, which occurred for one meal. These results further support the conclusion that this coding system is a highly reliable means for analyzing the number of food items using a visual food record.

Considering the assigned labels’ content, the number of assigned labels to each of the 19 GFFS categories did not vary significantly between the coders. This indicates a high agreement of their understanding and classification of the depicted food items. One exception was “Category C (Cereals and rice)” to which the coders assigned a significantly different number of meals. One explanation might be that it was challenging to distinguish food items from this category from those belonging to other categories and that one coder was better at handling this than the other. Perhaps also, food items belonging to category C were often difficult to recognize because there were components of mixed meals. It is also possible that one coder had a more advanced ability to identify single components in composite meals. Despite the coders’ limited agreement in this category, the coding system seems to enable a highly reliable coding of the content of the depicted food.

In terms of the coding systems’ reliability for the number of assigned food labels, intercoder reliability ranged from $\text{Kalpha} = .55$ to $.92$. This indicated that the minimum agreement between the coders for the number of assigned food labels was $55\%$ for one participant and that the maximum intercoder agreement was $92\%$ for another participant. The intercoder reliability for the remaining eight participants fell between these extremes. According to Krippendorff (2013),
the coding system’s reliability for the number of assigned food labels was high for some participants but medium or low for others.

In consideration of the total number of assigned food labels, both coders assigned the maximum and minimum number of labels to the same participants, respectively. Moreover, the difference in the number of assigned labels between coder A and B was not significant for any participant. Another result showed the percentage to which the coders assigned an identical or similar number of food labels as high for some participants, but as low for others.

Since the coding system’s overall reliability for the number of assigned labels was high, it seems appropriate for studies investigating the number of depicted food items for all participants’ meals. For example, the coding system could be employed in a study comparing the number of consumed food items between a sample suffering from an eating disorder and a sample without an eating disorder.

On the whole, the results for the coding system’s reliability for the number of assigned food labels per participants were mixed. Thus, the coding system does not appear to be ready for application in research concentrating on the number of consumed food items on an individual level. For example, the coding system should not be used in studies providing individual feedback to participants on the number of food items they ate.

4.2.2 Future research on the coding system’s reliability for food labels

Although the coding system showed a high reliability for assessing what people ate on an overall level, the system’s further development is desirable with emphasis on improving the reliability on an individual level.

For approximately 20% of all participants’ meals, the coders’ agreement was limited. Hence, a future study to identify reasons for the low reliability in the number of assigned labels for
some participants is needed. It might be that the coders’ understanding of the number of assigned labels different as one coder found three food items in a food image, whereas the other coder labeled five food items. Another cause for the difference in the number of assigned labels could be the procedure use of OptiDiet©. For example, a visual food record depicting Spaghetti Bolognese can be labeled as A) “Spaghetti Bolognese” (1 label) or as B) “Spaghetti”, “Bolognese” (2 labels). Consequently, despite a common understanding of the presented food as spaghetti bolognese, the coders could have assigned a different number of labels due to the range of labels available in OptiDiet©. Thus, clear coding must be established for such composite dishes. Further, it would be desirable to test the coders’ understanding of the coding rules within a training setting. Another way to address this issue would be to modify the large choice of food labels provided by OptiDiet©. With this objective, it would be beneficial to contact the OptiDiet© developers for recommendations.

Another important issue for future studies is to define the level of precision required. For example, the same food image could be coded as 1) “Spaghetti”, “Bolognese” (2 labels) or 2) “Spaghetti”, “Bolognese”, “Basil”, “Parmesan” (4 labels), artificially producing a different number of labels. Instructing, training, and testing the coders to work to a clearly defined level of precision could improve the coding system’s reliability for the number of assigned labels.

4.3 The coding system’s reliability for food nutrients

As well as how much and what people ate, the coding system’s reliability for the food nutrients on an overall level and on the level of each micro- and macro nutrient is discussed. First ideas for additional studies are provided.
4.3.1 Reliability for all and per nutrient

The coding system revealed an overall intercoder reliability of $K_{alpha} = .67$ for the food nutrients. This indicates a 67% agreement in the amount of all nutrients calculated with the nutrition software OptiDiet© based on the coder’s food weight estimation for all participants’ meals. In line with Krippendorff (2013), this reliability is low to medium. However, in light of the research field, an analysis of nutrients presents a sensitive task on a micro analysis level of dietary assessment. On the one hand, the found reliability is limited. On the other hand, it seems uncertain how reliably food nutrients can be determined at all when not directly measured.

Considering the coding system’s reliability for the single micro and macro nutrients, the intercoder reliability ranged from $K_{alpha} = .56$ to .79. That means the minimum agreement was 56% and the maximum agreement 79% at the level of single nutrients. The intercoder reliability for the other six micro- and macro nutrients fell between these extremes. Using Krippendorff’s (2013) guidelines, the coding system’s reliability for some nutrients was high and for others it was medium or low.

Because the coding system’s overall reliability for food nutrients was low to medium, it is questionable whether it is suitable for studies on the total amount of micro- and macronutrients. For instance, it cannot be recommended to use the coding system to compare the overall nutrients for a clinical population with an eating disorder and a sample without an eating disorder.

Since the results for the coding system’s reliability single food nutrients were mixed, it appears to be ready for use in studies analyzing certain food nutrients, but not others. For instance, the coding system could be employed to determine the amount of water contained in the participants’ meals, but not for calculating the fat contained in the meals.
4.3.2 Future research on the coding system’s reliability for food nutrients

Since the coding system’s reliability was not high for some food nutrients, further development is required. As observed by the researcher, sensitivity to changes in the food weight is different for food’s contained nutrients. A marginal change in the food weight for some items causes a large change for some nutrients. To determine the sensitivity indices of the 157 different nutrients calculated by OptiDiet© would be worthwhile since knowledge about nutrients’ sensitivity to changes in weight estimation could be used to develop reliable food weight estimation in studies focused on the more highly sensitive nutrients.

4.4 Limitations

To the best of my knowledge, this study was the first one that tested the reliability of the coding system of the smartphone based visual food record for Germany developed by Renner et al. (2015). Thereby it comprises a considerable contribution to the evaluation of a favorable up-to-date dietary assessment method. However, this study suffers from several limitations considering 1) excluded cases, 2) participants, 3) coders, and 4) Kalpha, all which are discussed below.

4.4.1 Excluded cases

In view of the excluded cases, the exclusion criteria and the handling of outliers can be interpreted critically.

Cases, i.e. meals, were primarily excluded if the coder failed to provide coding. However, this study also excluded beverages due to the focus on meals. A further exclusion criteria was established for food label content analysis. Only food items labeled with a GFFS code were included. Despite these guidelines, one inconsistency occurred. One coder did not provide a GFFS
code for one meal. In accordance with exclusion criteria, this meal could have been discounted for the content analysis but included in the analyses for food weight, number of assigned labels, and nutrients as the required data were available. However, this meal was excluded for all aspects of analysis.

Considering outliers, no outlier was excluded from the whole study. That means that meals were included in the calculations even if one or both coders provided an extremely high or extremely low food weight or number of food labels. Thus, these outliers could have distorted intercoder reliability through over- or underestimation.

In case of the repetition of this study, it would be beneficial to reconsider which meals should be excluded from which analysis through careful consideration of the exclusion criteria. In view of outliers, it seems important to study the reasons for extreme codings rather than simply excluding them. Did perhaps a very high number of assigned labels emerge because one coder labeled the meal very precisely while the other did not?

4.4.2 Participants

In light of the analyzed participants, number as well as sampling can be evaluated critically.

As only 10 participants from Renner et al.’s (2015) *SmartFood* study were analyzed to test the coding system’s intercoder reliability, the sample size can be regarded as small and its representativeness as questionable. Hence, the generalizability of the results is limited.

Moreover, the 10 participants were not chosen according to a proper randomization, but arbitrarily by one coder. Due to the lack of randomization, the analyzed participants could have certain characteristics that systematically influence their food intake regarding how much and what they ate which in turn could have distorted the coding system’s reliability.
Future research should include a larger and randomly chosen sample to provide more robust evidence on the coding system’s reliability.

4.4.3 Coders

In light of the two persons who coded the participants’ smartphone based visual food record, their randomization and blindness can be seen critically.

The coders were not drawn randomly. Rather, two employees of the Department of Health Psychology were selected, namely a PhD student and this thesis author. According to this lack of randomization the coders’ features, such as their interest in Health Psychology, could have systematically influenced the results. Thus, intercoder reliability might not only reflect the coding system’s quality but also the coders’ own features. For example, the coders’ interest and knowledge in health psychology could have influenced their coding.

Moreover, both coders were familiar with the study’s research questions. It is also conceivable that both coders desired the coding system to be a reliable tool for analyzing participants’ food intake. Presumably, the coders were highly motivated to do the coding as precisely as possible. Hence, it is imaginable that intercoder reliability not only results from the coding system’s quality but also from the coders’ ambition.

Repeating the study with randomized and blind coders could be preclude the above mentioned, potential biases. One option to achieve this would be to hire external coders who are unfamiliar with the study’s goals. In this context, it would be worthwhile to further study human features that influence food weight estimation and food labeling. For example, state variables such as motivation and concentration or trait variables such as conscientiousness and spatial awareness could be tested for their influence on food weight estimation and labeling meals.
4.4.4 Kalpha

In consideration of Kalpha, the metric chosen in the applied macro is discussed. Kalpha’s practicability also poses a problem due to software limitations. Furthermore, calculating Kalpha’s confidence levels and counting its arithmetic mean presents two controversial aspects.

The Kalpha macro was set on “ratio” as scale level for all calculations considering the variables 1) food weight, 2) number of food labels, and 3) food nutrients. For the variable number of food labels that seems appropriate. However, in dependence of the research question, it is debatable whether the variable food weight reached interval scale level since, considering that 100g French fries are not equal to 100g of salad. In this case, the equal distance of 100g food weight has different meaning. The same is true for nutrients. The distance of 100g can have a small meaning within a normal range and another meaning beyond a cut-off point, e.g. vitamin deficiency.

In addition to the chosen scale level for which Kalpha was calculated, Kalpha’s practicability can be seen critically. To calculate Kalpha, a separate SPSS data sheet had to be created to ensure that the data are listed side by side for comparison. For instance, to determine Kalpha for each nutrient, 157 files were created with 269 values of food weight estimation per meal. Due to the nature of the OptiDiet© output, it was not possible to copy and paste calculated nutrients. Consequently the nutrient value for each of the 269 meals for all 157 nutrient files had to be inserted manually and controlled for both coders (84,466 values). This process was time consuming and therefore undesirable.

Furthermore, adopting the macro to the study’s properties was burdensome. Theoretically, the macro’s parameters can easily be set to the data’s properties, but the macro did not initially run. Consequently, several weeks of intense email correspondence with the Kalpha inventor Klaus Krippendorff and the Macro developer Andrew Hayes were required. Eventually, Hayes rewrote the macro so that it would run with the study’s data. Thus, it is only with the support of
Krippendorff and Hayes that Kalpha could be calculated for this study. However, the feasibility of the Kalpha macro appears to be improving.

Another issue with the Kalpha macro was that the computer crashed on a regular basis, presumably as a result of the application of the macro to the study’s large dataset. As commonly used SPSS commands and the application of the macro to a smaller dataset did not cause the notebook to crash large scale data analysis with Kalpha is currently burdensome.

The calculation of Kalpha’s confidence levels created another issue. It was not possible to determine confidence levels for any of the calculated Kalphas with the macro. This was due to the fact that the macro implements a bootstrapping to determine confidence level, which was not possible based on the study’s dataset. In fact, the macro’s bootstrapping was insufficient to deal with the study’s large data range. Again, Krippendorff and Hayes were contacted. They have yet to find a solution for calculating the confidence levels for the study’s dataset for the macro. Krippendorff recommended using Kalpha as a descriptive measure. However, as an alternative, the Kalpha confidence levels were calculated with the standard SPSS setting. More precisely, the confidence levels could be calculated for the mean Kalphas, but not for the single Kalphas. Although, this seems currently to be the only solution to obtaining confidence levels, it is uncertain if this is a correct way of determining confidence levels for Kalpha.

Furthermore, calculating the arithmetic mean of Kalpha is controversial. The mean Kalpha was calculated for food weight, food labels, and food nutrients. For example, for the food weight’s mean Kalpha, the single food weight Kalphas from all 10 participants’ meals were added and the arithmetic mean was determined. However, when adding the calculated Kalphas for food weight estimation of all 269 meals and calculating their arithmetic mean, the mean Kalpha for food weight estimation differed by approximately .02 compared with the previously calculated mean Kalpha. This was the same for the number of assigned labels. This could be due to the mathematical
operations of Kalpha when calculating. Hence, it is questionable whether Kalpha is a reliable index to draw an arithmetical mean from. In the results section the lower Kalpha was presented to provide a more conservative measure and thereby avoid an overestimation of the coding system’s reliability.

In summary, it is important to consider the variables’ metric, to clarify their scale level. With the appropriate scale level the macro could be updated for a repetition of the study. The ease of application of the Kalpha macro was limited in this study. Thus, future research using Kalpha should be conducted with the awareness that calculating Kalpha can be time-consuming. Finally, it would be ideal to discuss issues such as methods for obtaining appropriate confidence levels and whether determining an arithmetic mean is appropriate with Krippendorff and Hayes.

4.5 Future research: Validation of the coding system

A major issue for future research is conducting a validity study for the coding system of Renner et al.’s (2015) smartphone based visual food record. This is of importance because intercoder reliability may be a result of coder similarity rather than the coding system’s quality. For example, Nelson et al. (1996) found gender, age, BMI and cognitive abilities as factors that potentially influence food weight estimation. The two coders of the present study were of the same gender and very similar considering age, BMI, and cognitive abilities. To clearly attribute the coder agreement to the coding system’s quality, coder similarity must be ruled out as explanation.

Table 9 summarizes the present study’s limitations and offers questions a validation study of the coding system should consider for their solution.
Table 9

<table>
<thead>
<tr>
<th>Limitations</th>
<th>Questions for future research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded cases</td>
<td><em>Exclusion criteria correctly implemented?</em>&lt;br&gt;<em>Exclude or analyze outliers?</em></td>
</tr>
<tr>
<td>Participants</td>
<td><em>Sufficient number of participants analyzed?</em>&lt;br&gt;<em>Participants sampled randomized?</em></td>
</tr>
<tr>
<td>Coders</td>
<td><em>Coders sampled randomized?</em>&lt;br&gt;<em>Coders blind to study?</em></td>
</tr>
<tr>
<td>Kalpha</td>
<td><em>Variables’ metric correctly set in the Kalpha macro?</em>&lt;br&gt;<em>Enough time to calculate Kalpha?</em>&lt;br&gt;<em>Possible to discuss with Hayes and Krippendorff how to obtain confidence levels for Kalpha?</em>&lt;br&gt;<em>Possible to discuss with Krippendorff and Hayes if an arithmetic mean of Kalpha is meaningful?</em></td>
</tr>
</tbody>
</table>

*Note.* Left: Criteria on which the present reliability study was limited. Right: Questions future research could consider to preclude these limitations.

For example, the coders should be randomly chosen and blind to the study’s goals. To determine the food weight’s reliability, an empirically founded food selection must first be directly weighed by the researchers. Thereafter, pictures of the food must be taken. Next, several coders should receive these pictures and estimate the weight of the depicted food. Thereby, a direct comparison between each coder’s food weight estimation and the real food weight could be made. To detect any significant difference, a t-test for independent samples could be used. Additionally, the Kalpha between the coders’ food weight estimations must be calculated. If no significant difference is found between their food weight estimation and the real food weight and if the Kalpha is high, the coding system can then be stated as allowing for a valid coding for the food weight estimation.

### 4.6 Conclusion

Table 10 sums up the reliability for food weight, food labels and food nutrients and provides ideas for future improvements.
Table 10

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reliability</th>
<th>Ideas for improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>For all participants</td>
<td>Per participant</td>
</tr>
<tr>
<td>Food weight</td>
<td>high</td>
<td>low-high</td>
</tr>
<tr>
<td>Food labels</td>
<td>high</td>
<td>low-high</td>
</tr>
<tr>
<td>Food nutrients</td>
<td>medium</td>
<td>low-high</td>
</tr>
</tbody>
</table>

Note. Left: Variables investigated in the present reliability study. Middle: Found reliability for all and per participant. Right: Ideas for future research to enhance reliability.

The highest intercoder reliability was found for the number of assigned labels. The reliability for food weight estimation was a little lower, and the reliability for food nutrients was the lowest. For both the food weight estimation and the food labeling, the reliability varied strongly between single participants. Reliability was excellent for some, medium for others, and was insufficient for a few. Moreover, reliability varied greatly for the 157 food nutrients.

In light of the research context, the coding system’s reliability appears high for food weight estimation and the number of food labels. However, the coding system does not seem ready to reliably determine the food nutrients. Overall, the coding system was appropriate on a macro level, for instance, for the mean reliability of food weight and food labeling.

Nonetheless, its reliability appears limited for close analysis on the participant level and its reliability seems even lower on the more detailed level of nutrient analysis. Thus, the coding system can provide highly reliable data for research questions considering overall food weight or food labeling. For research on a participant or nutrient level, the coding system could be further developed to allow for reliable conclusions.

Extension of the food atlas is desirable to improve the food weight estimation. This study did add to the sample of food images, but further expansion of the selection is desirable. Considering food labels, a clear formulation of coding rules regarding the choice of labels for cases
where different labels can be applied to the same item could improve reliability. To enhance the coding system’s reliability at the level of nutrient, a study investigating nutrients’ sensitivity to food weight estimation could provide promising results. Moreover, it would be interesting to compare reliability rankings for food weight and food labels between participants. Does a participant with a low reliability in food weight estimation also has a low reliability in the number of assigned food labels? Studies on relations like these could be informative for the further development of the coding system.

This study was the first one testing the coding system’s reliability for a smartphone based visual food record for Germany. Therefore, it contributed to the development of an appropriate and up-to-date dietary assessment tool for Germany. Furthermore, this study revealed the limitations of the coding system and offered ideas for further research.

Approximately 20 years ago Nelson et al. (1996) stated, “There is no single measure of habitual diet in free-living subjects which is entirely valid” (p. 32). Although this view may be still true today, a rapid and great progress in the development and evaluation of dietary assessment methods for measuring people’s food intake under real life conditions can be observed. By evaluating the coding system’s reliability and extending the food atlas this study contributed to testing and improving the coding system’s psychometric quality. Having a comprehensive smartphone based visual food record with a highly reliable and valid coding system can open new opportunities for clinical and health psychology. This study contributed to realize that vision.
References


References


Kolodziejczyk, J. K., Merchant, G., & Norman, G. J. (2012). Reliability and validity of child/adolescent food frequency questionnaires that assess foods and/or food groups. *Journal of Pediatric Gastroenterology and Nutrition, 55*(1), 4-13. doi:10.1097/MPG.0b013e318251550e


Vereecken, C., Dohogne, S., Covents, M., & Maes, L. (2010). How accurate are adolescents in portion-size estimation using the computer tool young adolescents’ nutrition assessment on computer (YANA-C)? *British Journal of Nutrition, 103*(12), 1844-1850. doi:10.1017/S0007114510000127


# Appendix

Table 4

<table>
<thead>
<tr>
<th>Participant</th>
<th>N meals</th>
<th>Mean of differences</th>
<th>Differences in food weight estimation</th>
<th>Absolute</th>
<th>Percent</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>72.43g</td>
<td>Same</td>
<td>0g</td>
<td></td>
<td>3</td>
<td>9.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 0g&lt;x&lt;10g</td>
<td>5</td>
<td>16.13</td>
<td>5.89</td>
<td>3.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 10g&lt;x&lt;40g</td>
<td>7</td>
<td>22.58</td>
<td>23.11</td>
<td>5.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 40g&lt;x&lt;80g</td>
<td>9</td>
<td>29.03</td>
<td>65.44</td>
<td>15.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 80g&lt;x&lt;125g</td>
<td>7</td>
<td>22.58</td>
<td>209.29</td>
<td>146.08</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>168.61g</td>
<td>Same</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 25g&lt;x&lt;55g</td>
<td>7</td>
<td>25</td>
<td>38</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 55g&lt;x&lt;104g</td>
<td>7</td>
<td>25</td>
<td>81</td>
<td>11.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 104g&lt;x&lt;261g</td>
<td>7</td>
<td>25</td>
<td>155.79</td>
<td>49.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 261g&lt;x&lt;865g</td>
<td>7</td>
<td>25</td>
<td>399.64</td>
<td>212.54</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>126.97g</td>
<td>Same</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 3g&lt;x&lt;40g</td>
<td>6</td>
<td>26.09</td>
<td>24.17</td>
<td>16.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 40g&lt;x&lt;85g</td>
<td>5</td>
<td>21.74</td>
<td>56.8</td>
<td>8.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 85g&lt;x&lt;175g</td>
<td>7</td>
<td>30.43</td>
<td>119.29</td>
<td>33.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 175g&lt;x&lt;542g</td>
<td>5</td>
<td>21.74</td>
<td>331.2</td>
<td>120.78</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>77.65g</td>
<td>Same</td>
<td>0g</td>
<td></td>
<td>1</td>
<td>4.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 0g&lt;x&lt;28.75g</td>
<td>4</td>
<td>18.18</td>
<td>14.25</td>
<td>10.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 28.75g&lt;x&lt;59.6g</td>
<td>6</td>
<td>27.27</td>
<td>38.7</td>
<td>10.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 59.6g&lt;x&lt;126.38g</td>
<td>6</td>
<td>27.27</td>
<td>90.67</td>
<td>27.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 126.38g&lt;x&lt;210g</td>
<td>5</td>
<td>22.72</td>
<td>166.25</td>
<td>14.54</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
<td>70.52g</td>
<td>Same</td>
<td>0g</td>
<td></td>
<td>3</td>
<td>10.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 0g&lt;x&lt;15g</td>
<td>5</td>
<td>17.24</td>
<td>11</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 15g&lt;x&lt;50g</td>
<td>6</td>
<td>20.69</td>
<td>34.17</td>
<td>5.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 50g&lt;x&lt;97.5g</td>
<td>8</td>
<td>27.58</td>
<td>66.88</td>
<td>19.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 97.5g&lt;x&lt;500g</td>
<td>7</td>
<td>24.14</td>
<td>178.57</td>
<td>143.11</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>91.44g</td>
<td>Same</td>
<td>0g</td>
<td></td>
<td>1</td>
<td>4.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 0g&lt;x&lt;30.5g</td>
<td>5</td>
<td>10.83</td>
<td>21</td>
<td>7.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 30.5g&lt;x&lt;68g</td>
<td>6</td>
<td>25.00</td>
<td>59.5</td>
<td>18.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 68g&lt;x&lt;139.5g</td>
<td>6</td>
<td>25.00</td>
<td>91.83</td>
<td>17.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 139.5g&lt;x&lt;255g</td>
<td>6</td>
<td>25.00</td>
<td>206.5</td>
<td>46.35</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>64.64g</td>
<td>Same</td>
<td>0g</td>
<td></td>
<td>2</td>
<td>9.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 0g&lt;x&lt;11.25g</td>
<td>3</td>
<td>14.29</td>
<td>6.33</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 11.25g&lt;x&lt;20g</td>
<td>3</td>
<td>14.29</td>
<td>15.5</td>
<td>3.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 20g&lt;x&lt;75g</td>
<td>8</td>
<td>38.10</td>
<td>41.50</td>
<td>21.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 75g&lt;x&lt;330g</td>
<td>5</td>
<td>23.81</td>
<td>192</td>
<td>102.08</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>105.48g</td>
<td>Same</td>
<td>0g</td>
<td></td>
<td>2</td>
<td>5.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 0g&lt;x&lt;20g</td>
<td>7</td>
<td>20</td>
<td>12.86</td>
<td>5.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 20g&lt;x&lt;55g</td>
<td>7</td>
<td>20</td>
<td>36.93</td>
<td>8.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 55g&lt;x&lt;130g</td>
<td>11</td>
<td>31.43</td>
<td>84.09</td>
<td>26.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 130g&lt;x&lt;914g</td>
<td>8</td>
<td>22.66</td>
<td>302.28</td>
<td>256.28</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>81.48g</td>
<td>Same</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 10g&lt;x&lt;35g</td>
<td>6</td>
<td>30</td>
<td>22</td>
<td>10.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 35g&lt;x&lt;70g</td>
<td>4</td>
<td>20</td>
<td>55</td>
<td>9.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 70g&lt;x&lt;134.5g</td>
<td>5</td>
<td>25</td>
<td>93.7</td>
<td>23.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 134.5g&lt;x&lt;173g</td>
<td>5</td>
<td>25</td>
<td>161.8</td>
<td>15.24</td>
</tr>
<tr>
<td>10</td>
<td>36</td>
<td>58.74g</td>
<td>Same</td>
<td>0g</td>
<td></td>
<td>4</td>
<td>11.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Percentile 0g&lt;x&lt;13g</td>
<td>5</td>
<td>13.89</td>
<td>8.2</td>
<td>2.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Percentile 13g&lt;x&lt;27.5g</td>
<td>9</td>
<td>25.00</td>
<td>21.5</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Percentile 27.5g&lt;x&lt;83.75g</td>
<td>9</td>
<td>25.00</td>
<td>49.78</td>
<td>18.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Percentile 83.75g&lt;x&lt;370g</td>
<td>9</td>
<td>25.00</td>
<td>159.11</td>
<td>88.03</td>
</tr>
</tbody>
</table>

Note. Table 4 shows the differences in food weight estimation between coder A and B per participant. The differences in food weight estimation are divided into percentiles. The case of same weight estimation is also depicted.