



# Reproducibility, power and validity of visual analogue scales in assessment of appetite sensations in single test meal studies

A Flint<sup>1\*</sup>, A Raben<sup>1</sup>, JE Blundell<sup>2</sup> and A Astrup<sup>1</sup>

<sup>1</sup>Research Department of Human Nutrition, Center for Food Research, The Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark; <sup>2</sup>BioPsychology Group, Department of Psychology, University of Leeds, Leeds, UK

**OBJECTIVE:** To examine reproducibility and validity of visual analogue scales (VAS) for measurement of appetite sensations, with and without a diet standardization prior to the test days.

**DESIGN:** On two different test days the subjects recorded their appetite sensations before breakfast and every 30 min during the 4.5 h postprandial period under exactly the same conditions.

**SUBJECTS:** 55 healthy men (age 25.6±0.6 y, BMI 22.6±0.3 kg/m<sup>2</sup>).

**MEASUREMENTS:** VAS were used to record hunger, satiety, fullness, prospective food consumption, desire to eat something fatty, salty, sweet or savoury, and palatability of the meals. Subsequently an *ad libitum* lunch was served and energy intake was recorded. Reproducibility was assessed by the coefficient of repeatability (CR) of fasting, mean 4.5 h and peak/nadir values.

**RESULTS:** CRs (range 20–61 mm) were larger for fasting and peak/nadir values compared with mean 4.5 h values. No parameter seemed to be improved by diet standardization. Using a paired design and a study power of 0.8, a difference of 10 mm on fasting and 5 mm on mean 4.5 h ratings can be detected with 18 subjects. When using desires to eat specific types of food or an unpaired design, more subjects are needed due to considerable variation. The best correlations of validity were found between 4.5 h mean VAS of the appetite parameters and subsequent energy intake ( $r=±0.50–0.53$ ,  $P<0.001$ ).

**CONCLUSION:** VAS scores are reliable for appetite research and do not seem to be influenced by prior diet standardization. However, consideration should be given to the specific parameters being measured, their sensitivity and study power.

*International Journal of Obesity* (2000) 24, 38–48

**Keywords:** visual analogue scales; hunger; satiety; diet standardization

## Introduction

Subjective sensations of hunger, satiety, other appetite sensations and desires to eat specific types of food may be influenced by a number of different internal factors, including physiological and psychological variables.<sup>1</sup> Moreover, external factors, such as prior meals, physical activity, temperature, weather etc. may influence the subjective sensations on the test day. When assessing the reproducibility, as many factors as possible should therefore be kept constant. Reproducibility refers to the variation between measurements made at different time points on different test days, but using the same standardized method.

In order to assess subjective appetite sensations, visual analogue scales (VAS) are often used. VAS are most often composed of lines (of varying length) with words anchored at each end, describing the extremes (that is, 'I have never been more hungry'/'I am not

hungry at all'). Subjects are asked to make a mark across the line corresponding to their feelings. Quantification of the measurement is done by measuring the distance from the left end of the line to the mark.

The reproducibility and validity of VAS scores has been quite well studied in other research areas, especially concerning pain.<sup>2–4</sup> Thus, in pain research VAS is used as 'the gold standard'.<sup>5</sup> The results from other research areas cannot, however, be extrapolated to appetite research. In this field it appears that only a very limited number of studies have investigated the reproducibility and validity of the VAS method.

Some studies report good test–retest results between identical or almost identical trials, using paired rank-sum test or correlations.<sup>6–10</sup> However, according to Bland and Altman<sup>11</sup> neither of these statistical methods are sufficient to describe reproducibility of a method. Using the method of Bland and Altman, Stratton *et al*<sup>12</sup> showed that reproducibility was good when ratings of appetite were made immediately one after another on the same test day. On the other hand, when tested twice in the same subjects on different days but under standardized pre-test conditions, Raben *et al*<sup>13</sup> found relatively large repeatability coefficients for appetite ratings after identical meals,

\*Correspondence: A Flint, Research Department of Human Nutrition, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.  
Email: afl@kvl.dk  
Received 18 December 1998; revised 23 June 1999; accepted 27 July 1999

thereby questioning the reproducibility of the VAS method.

The validity of a method refers to the ability of a method to measure what it is intended to do. The overall picture from the literature is that VAS ratings of 'hunger', 'desire to eat', 'appetite for a meal', or 'prospective food consumption' before a meal to a certain degree is related to the subsequent food intake,<sup>6,9,10,14–16</sup> while ratings of 'satiety' or 'fullness' have been found not to relate to the forthcoming food intake.<sup>10,16</sup>

As stated by de Graaf<sup>17</sup> the relationship between appetite ratings and energy intake can be calculated in several ways. Depending on the type of correlation which is calculated, the magnitude of the correlation coefficient and thereby the conclusions made can be very different using the same data material.

Thus, although widely used since the early 1960s, the reproducibility and validity of appetite scores in studies involving test meals has not been clarified. Furthermore, no study has investigated the effect of diet standardization of subjects prior to the test days. In the present study the questions of reproducibility, short term validity and power of appetite scores were addressed. Conditions with or without previous diet standardization were also compared.

## Subjects and methods

The study was carried out at the Research Department of Human Nutrition, Centre for Food Research, The Royal Veterinary and Agricultural University, Frederiksberg, Denmark, and was approved by the Municipal Ethical Committee of Copenhagen and Frederiksberg as in accordance with the Helsinki-II declaration.

Appetite and palatability ratings were assessed twice in the same subject before and after identical breakfast meals on two different occasions, 32 subjects with and 23 subjects without a diet standardization prior to the test days. *Ad libitum* lunch was served 4.5 h after the breakfast and energy intake was registered.

### Subjects

Fifty-five healthy, normal-weight male subjects, 19–36 y of age participated in the study. None were smokers, elite-athletes, or had a history of obesity or diabetes. One group (non-diet group, ND,  $n = 23$ ) was instructed to abstain from physical activity and alcohol for 2 days prior to each test day. On the test day the subjects arrived at the department having used the least strenuous transport possible and having fasted for 12 h. The other group (diet group, D,  $n = 32$ ) followed the same instructions and consumed a standard diet for 2 days prior to the test days. The energy content was estimated according to each subject's

individual energy requirements, determined from WHO tables according to age, weight, height and sex.<sup>18</sup> The multiplication factor 1.78 was used corresponding to medium physical activity. The subjects of the non-diet group (ND) were  $26.4 \pm 1.0$  (mean  $\pm$  s.e.m.) years of age, and had a body mass index (BMI) of  $22.5 \pm 0.5$  kg/m<sup>2</sup>, while the diet group (D) were  $25.0 \pm 0.6$  y of age, and had a BMI of  $22.7 \pm 0.4$  kg/m<sup>2</sup> (ns). Their average energy requirements were  $14.0 \pm 0.2$  MJ/d.

All subjects gave written consent after the experimental protocol had been explained to them.

### Design

All subjects were tested on two different occasions separated by 1–3 weeks. On the two test days the subjects arrived at the department at 9:00 h. with a minimum of physical activity using car, bus or train. The subjects' weight and height were measured. At 9:40 h the first questionnaire on appetite was filled in and the breakfast was served at 9:45 h. From 10:00 h to 14:00 h the subjects filled in questionnaires every 30 min. At 14:15 h lunch was served, and the subjects could eat *ad libitum* until 'comfortable satisfaction'. During the day the subjects could sit in our experimental dining room, read, walk quietly around (also for a short while outside the dining room), listen to the radio or watch TV/video (light entertainment). They were allowed to consume water if necessary and toilet visits. The exact hour, time spent, type of activity, and the amount of water consumed on the first test day were noted and repeated on the second test day. On a test day two to eight subjects were tested, all seated in the same dining room. They could talk with each other as long as the conversation did not involve food, appetite and related issues.

### Diets

Two different test meals were used, one for breakfast and one for lunch. The breakfast consisted of yoghurt, bread, butter, cheese, jam, kiwi-fruit, orange juice and water. Total available energy content of the meal was 2 MJ and the energy composition was similar to the standard diet (50 E% carbohydrate, 37 E% fat, 13 E% protein, 2.2 g/MJ dietary fibre). The lunch was a homogenous mixed hot-pot consisting of pasta, minced meat, green pepper, carrots, squash, onions, corn and cream from which the subjects could eat *ad libitum* (50 E% carbohydrate, 37 E% fat, 13 E% protein, 0.7 g/MJ dietary fibre).

### Questionnaires

VAS, 100 mm in length with words anchored at each end, expressing the most positive and the most negative rating, were used to assess hunger, satiety, fullness, prospective food consumption, desire to eat something fatty, salty, sweet or savoury (Appendix 1), and the palatability (five questions) of the test meal

(Appendix 2). The questionnaires were made as small booklets showing one question at a time. Subjects did not discuss or compare their ratings with each other and could not refer to their previous ratings when marking the VAS.

### Statistical analyses

Palatability of the test meals and fasting values, 4.5 h means and peaks/nadirs of the appetite ratings and ratings of specific desires were compared by paired *t*-tests. The postprandial response curves of the appetite ratings and rating of specific desires were compared by parametric analysis of variance (ANOVA) using an epsilon-corrected split-plot analysis with time as a factor.

Tests of reproducibility were performed according to Bland and Altman.<sup>11</sup> Thus, the coefficient of repeatability (CR = 2s.d.) on the mean differences between meal 1 and meal 2 was calculated for fasting values, 4.5 h means, and peaks/nadirs of the appetite ratings, for fasting and 4.5 h mean ratings of specific desires and for the palatability ratings.

Power calculations were made according to the method of Altman<sup>19</sup> using a nomogram. Validity was judged by correlation analysis between VAS ratings (prelunch values, pre–post difference and postprandial 4.5 h mean ratings on the one hand) and energy intake at the *ad libitum* lunch, on the other hand.

For all statistical analysis the Statistical Analysis Package (SAS Institute, Cary, NC, version 6.08 and 6.11) was used. The coefficient of variation (CV) was calculated for 4.5 h mean as:

$$CV = \frac{\sqrt{\sum(\text{test1} - \text{test2})^2 / (2 \times \text{pairs})}}{\text{mean}}$$

## Results

### Reproducibility

**Appetite scores.** The response curves for satiety, hunger, fullness and prospective food consumption are presented in Figure 1 for both groups. The profiles were similar after the two test meals and no statistically significant differences between test days were found by ANOVA.

Table 1 shows the results for reproducibility of fasting values, mean ratings and peak/nadir. Considering fasting ratings, none of the parameters differed between test days, and CR ranged from 24 to 30 mm in the ND group and from 23 to 30 mm in the D group for the four different appetite scores. Mean 4.5 h appetite scores showed no differences in any parameter between test days in the ND group, while

subjects felt they could eat less on the second test day in the D group (diet effect:  $P = 0.021$ , Figure 2). CRs ranged from 17 to 24 mm in the ND group and from 15 to 17 mm in the D group. For illustration, Figure 3 shows a Bland–Altman plot for mean values of hunger in the D group, showing the CR to be 16 mm. The coefficients of variation were 9–21% and 7–24% for 4.5 h means for the ND and D groups, respectively, for appetite sensations, with no significant differences between the groups. However, a borderline significant lower CV value was found in the case of 4.5 h mean hunger in the D group compared with the ND group ( $F = 2.14$ , d.f. = 22, 31,  $P = 0.05$ ).

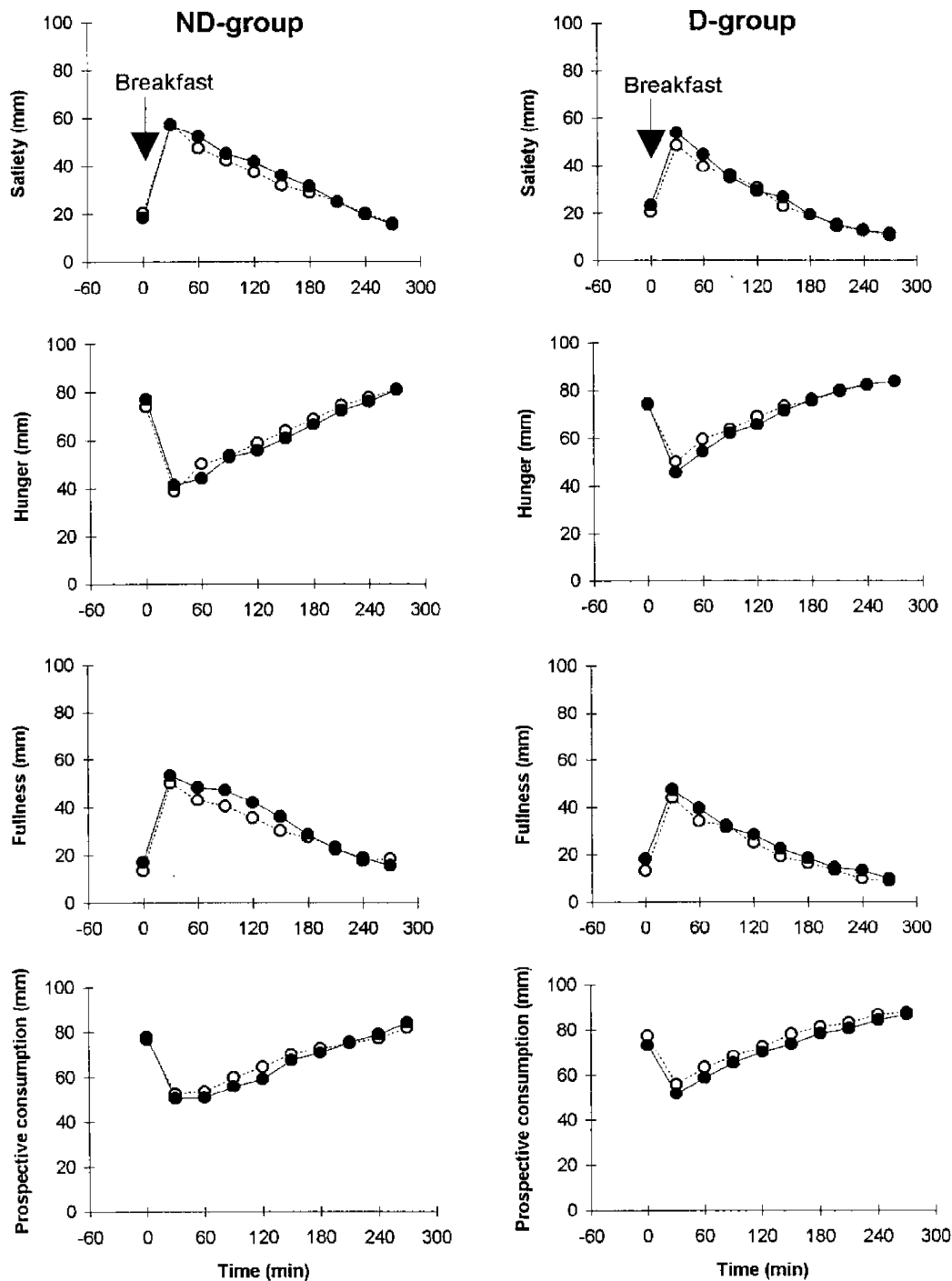
Correlations between ratings on the first and second test day varied considerably, being weakest for fasting values and strongest for 4.5 h mean values (Table 1).

**Desire for specific food types.** The response curves for the desire of different food types are shown in Figure 4 for both groups. No significant differences were found, except that the profile curves of desire for something sweet were significantly different between the two test days ( $P < 0.0001$ ; more flat on the first test day). Table 2 and Figure 5 show the results for fasting values and mean ratings. No differences were found in fasting ratings between test days, except in the case of desire for something savoury. The subjects had a greater desire to eat something savoury ( $P = 0.014$ ) on the second day in the ND group. CRs ranged from 29 to 61 mm in the ND group and from 28 to 40 mm in the D group.

Mean 4.5 h scores of desires showed no differences in any parameter between test days in the ND group, while subjects felt less desire to eat something savoury on the second day ( $P = 0.05$ , Figure 5). No differences were found between groups with regard to variances, CVs being 12–20% and 13–28% respectively for the ND and D group for 4.5 h means.

Strong and significant correlations were seen in all cases except for desire to eat something sweet in the fasting state in the ND group (Table 2).

**Palatability of the meals.** Ratings for visual appeal, taste, smell, aftertaste and overall palatability after the two breakfasts and lunches are shown in Figure 6. Apparently the ratings were very similar after equal meals on the two test days. However, in the ND group the taste of the breakfast was recorded to be best the first time ( $P = 0.005$ ), while all other palatability scores did not differ. The CRs for all five questions were quite large for both breakfast and lunch (Table 3), ranging from 20 mm (taste) to 47 mm (visual appeal). In the D group no differences between meals were observed in any of the palatability scores, CRs ranging from 28 mm (visual appeal) to 37 mm (aftertaste).



**Figure 1** Subjective appetite scores after two identical test meals in the ND group ( $n=23$ ) and the D group ( $n=32$ ) on two test days (test 1: open circles; test 2: filled circles). Values are means. The arrow indicates the initiation of the test meal (breakfast). No differences by ANOVA.

Correlation coefficients varied between 0.42 and 0.85 for relations between ratings of test days 1 and 2 (Table 3), all being significant.

**Power calculations**

Since no effect of prior diet standardization was seen, except for a borderline significant difference in the

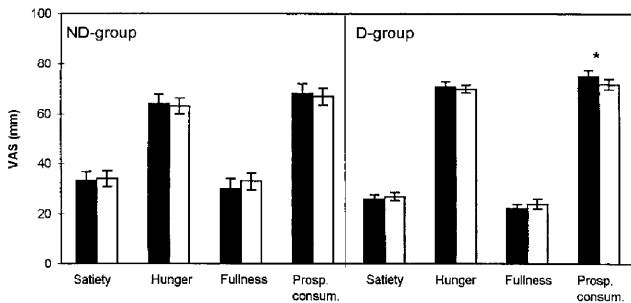
case of hunger, the data from the two groups were pooled.

In Table 4 the number of subjects needed to detect a difference in VAS scores in a paired and in an unpaired design is shown. In the case of a paired design, calculations were made at two levels of detectable differences (5 and 10 mm) and at two levels of power (0.8 and 0.9 corresponding to a 20% or 10% chance of type 2 error, respectively). In the

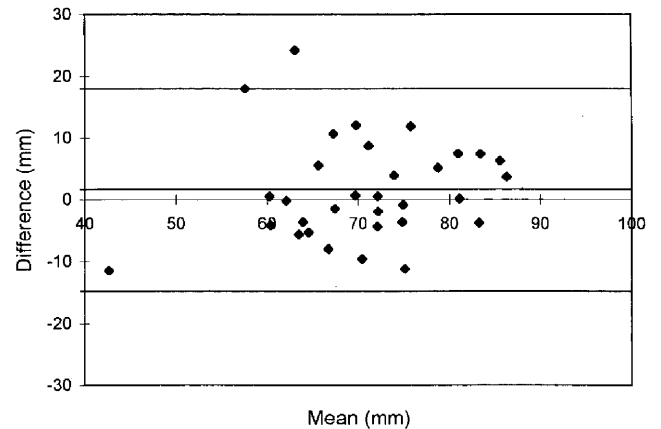
**Table 1** Reproducibility of appetite scores

	ND group			D group		
	Mean difference (mm)	CR (mm)	r	Mean difference (mm)	CR (mm)	r
<i>Fasting</i>						
Satiety	2	30	0.17	-3	29	0.45**
Hunger	-3	28	0.26	0	23	0.51**
Fullness	-3	27	0.16	-5	30	0.23
Prospective food consumption	-1	24	0.47*	4	29	0.47**
<i>4.5 h mean</i>						
Satiety	-1	20	0.73***	-2	17	0.61***
Hunger	1	24	0.61**	2	16	0.68***
Fullness	-3	18	0.81***	-3	16	0.76***
Prospective food consumption	1	17	0.87***	3*	15	0.85***
<i>Peak</i>						
Satiety	1	33	0.61**	-5	34	0.44*
Fullness	-6	41	0.43*	-4	38	0.57***
<i>Nadir</i>						
Hunger	-3	41	0.54**	4	39	0.40*
Prospective food consumption	2	26	0.81***	4	24	0.80***

ND group: subjects not diet standardized prior to test days ( $n=23$ ). D group: subjects diet standardized 2 days prior to test days ( $n=32$ ). Mean difference: mean of (test 1 - test 2). \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .



**Figure 2** Subjective 4.5 h mean appetite scores on two occasions (test 1: dark bars; test 2: light bars). Values are means  $\pm$  s.e.m. \* $P < 0.05$  by paired  $t$ -test.



**Figure 3** Bland-Altman diagram of 4.5 h mean hunger scores from the D group. Mean values (mean of test 1 and test 2) are plotted against the difference of the same two ratings. The interval  $\text{mean} \pm 2 \text{ s.d. (mean} \pm \text{CR)}$  is shown by horizontal lines.

case of an unpaired design only calculations with a study power of 0.8 are shown because a higher study power results in an unrealistically high number of subjects for most test meal studies. Choosing a power of 0.8 and a paired design, 70 subjects would be needed to detect a difference of 5 mm in fasting values and 18 subjects to detect a difference of 10 mm with regard to appetite sensations. Looking at 4.5 h mean values the number of subjects is considerably lower, while peak/nadir values require the largest number of subjects. More subjects are needed to assess the desire to eat specific types of food due to a larger variation in these data.

Due to the component of between-subject variation more subjects are needed when using the unpaired design compared to the paired design (Table 4).

**Validity**

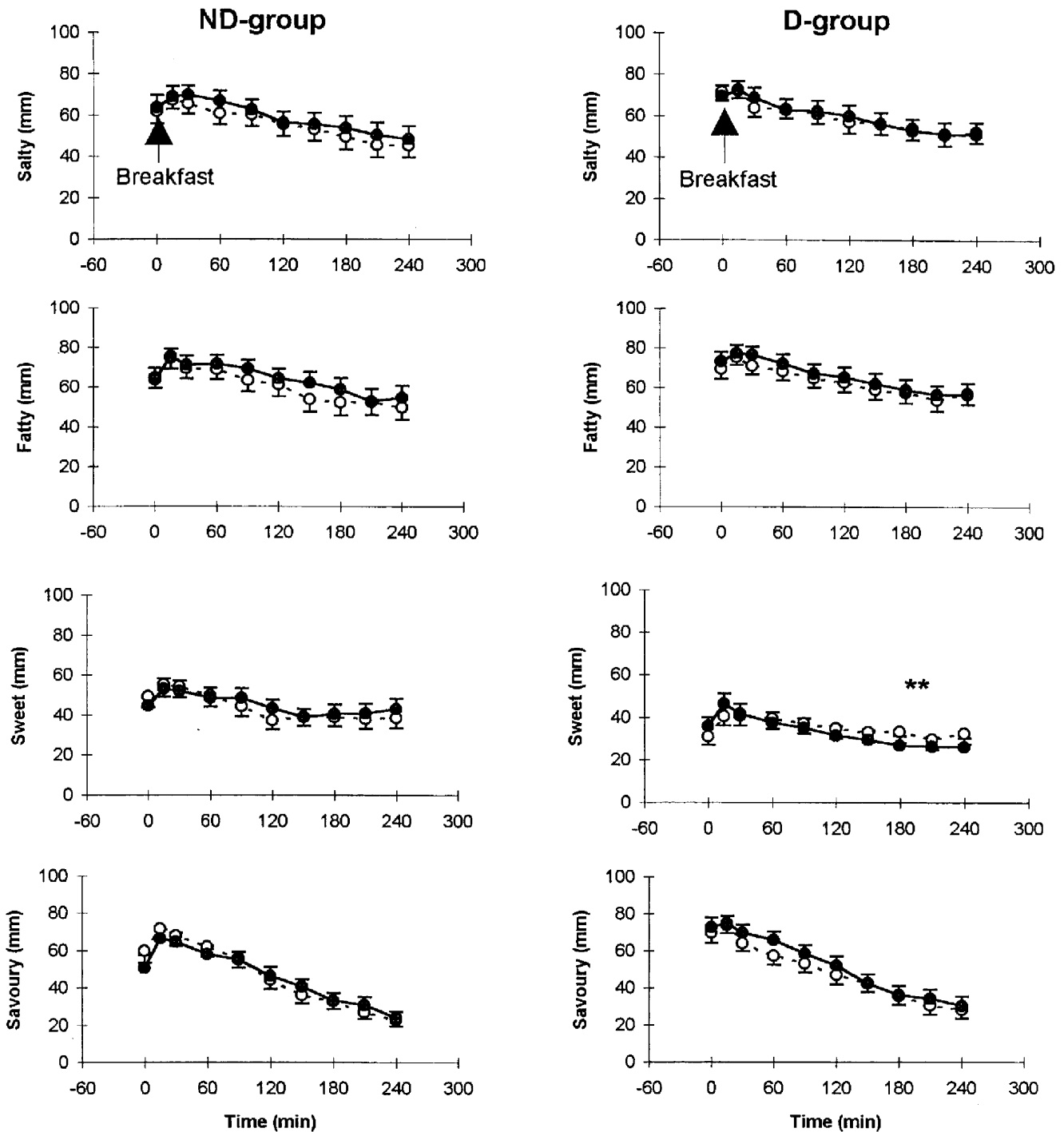
No difference was seen between *ad libitum* energy intake for the lunch meal on test days 1 and 2. Average energy intake was  $5.3 \pm 1.4$  MJ on the first day and  $5.2 \pm 1.4$  MJ on the second day.

Correlation coefficients between VAS scores and the *ad libitum* energy intake are shown in Table 5.

Prelunch values, the difference between pre- and postlunch values and the 4.5 h mean value of postprandial values after the breakfast were correlated to the subsequent lunch energy intake (EI). Data from both groups were pooled.

As indicated in Table 5, an outlier in the hunger rating immediately before lunch on day 1 was left out of the analysis of hunger. The rating was 3.3 s.d. away from the mean, and furthermore the rating was not in line with the rest of that subject's own ratings. However, the analysis was made with and without the outlier. When the outlier was included, the correlation coefficient for pre-lunch values vs EI was 0.16,  $P = 0.257$ . The outlier did not influence the analysis of postprandial 4.5 h mean in relation to EI.

Postprandial 4.5 h mean VAS ratings were more strongly correlated to energy intake compared with pre-lunch and pre-post difference ratings. Stronger correlations were observed on test day 2 compared



**Figure 4** Subjective scores for desire for specific types of food after two identical test meals in the ND group ( $n=23$ ) and the D group ( $n=32$ ) on two test days (test 1: open circles; test 2: filled circles). Values are means. The arrow indicates the initiation of the test meal. No differences by ANOVA, except in the case of desire for something sweet (\*\* $P < 0.01$ ).

with test day 1 or by using mean values of test days 1 and 2 instead of data from either test day 1 or 2.

## Discussion

### Reproducibility

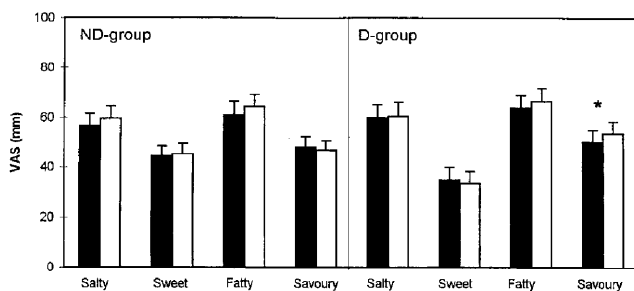
The temporal trackings of mean ratings of appetite and desires for specific foods show similar profiles.

However, in the underlying data considerable variation in observations is present. This variation is the sum of true biological day-to-day variation and methodological variation. There is at present no way to distinguish between these two types of variation. The relatively large variation is not surprising since we are dealing with subjective feelings. Using the 100 mm VAS for assessment of appetite and desires of food types we find that the reproducibility is relatively low,

**Table 2** Reproducibility of desires for specific foods

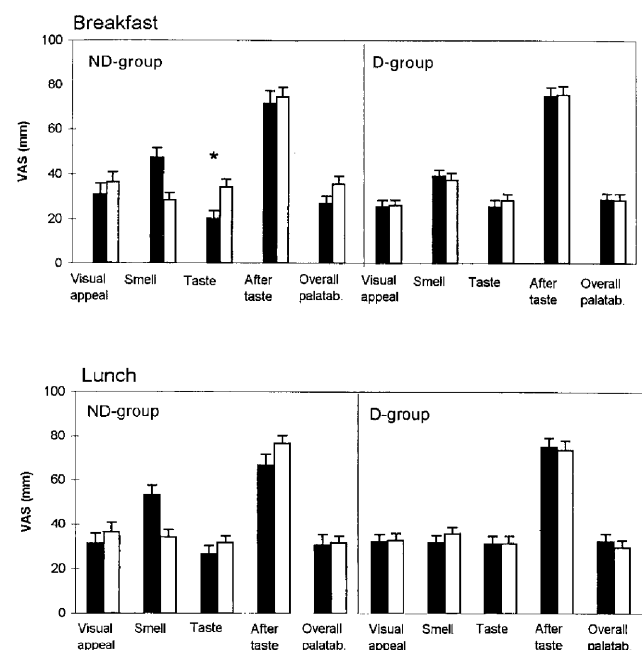
	ND group			D group		
	Mean difference (mm)	CR (mm)	r	Mean difference (mm)	CR (mm)	r
<i>Fasting</i>						
Salty	-2	29	0.87***	2	40	0.72***
Sweet	5	61	0.15	-5	28	0.81***
Savoury	9*	32	0.87***	-3	29	0.89***
Fatty	1	39	0.74***	-4	35	0.79***
<i>4.5 h mean</i>						
Salty	-3	27	0.85***	0	28	0.86***
Sweet	1	23	0.83***	2	27	0.83***
Savoury	1	27	0.76***	-4*	19	0.91***
Fatty	-3	21	0.91***	-3	24	0.88***

ND group: subjects not diet standardized prior to test days ( $n=23$ ). D group: subjects diet standardized 2 days prior to test days ( $n=32$ ). Mean difference: mean of (test 1 - test 2). \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .



**Figure 5** Subjective 4.5 h mean scores for desire for specific types of food on two occasions (test 1: dark bars; test 2: light bars). Values are means  $\pm$  s.e.m. \* $P < 0.05$  by paired  $t$ -test.

when expressed as fasting, mean or peak/nadir ratings, compared with objective methods like measurements of energy expenditure and substrates in the blood—parameters often used in combination with investigations of appetite. The highest degree of reproducibility was found for mean 4.5 h appetite values and the lowest degree of reproducibility for fasting ratings for desires of specific food items. This high degree of reproducibility of mean values is not surprising, since the role of a single outlying or erroneous rating will be reduced. In addition, the anticipation of the next meal probably influenced the ratings to converge toward one end of the appetite scale as lunch-time approached. Stratton *et al*<sup>12</sup> found coefficients of repeatability (CRs) between 1.6 and 12.8 mm, when ratings were made immediately one after another on the same test day. However, to our knowledge the study by Raben *et al*<sup>13</sup> is the only one which has calculated CRs between appetite ratings on two separate days, using the same meal. They found CRs to be very similar to our CRs for postprandial mean values, while this study showed lower CRs for fasting (24–30 vs 29–52 mm) and peak/nadir values (24–41 vs 32–54 mm). The explanation may be the different number of subjects included in the studies (23 and 32 vs 9). Having a mean value of 50 mm and a CR of 24 mm means that, with no change in the experimental setting, a second rating under identical conditions a week later may range from 26 to 74 mm, covering half of the total range of the scale.



**Figure 6** Palatability scores of identical test meals on two occasions (test 1: dark bars; test 2: light bars). Values are means  $\pm$  s.e.m. \* $P < 0.05$  by paired  $t$ -test.

The diet standardization prior to the test days did not improve the reproducibility, apart from a slightly better 4.5 h mean rating of hunger ( $P=0.05$ ). Therefore, appetite ratings done by VAS are seemingly not sensitive to this intervention. However, the difference between the standard diet and the habitual diet of this group may not have been great. Including other groups of subjects, e.g. restrained eaters, who might be influenced by taking part in a diet-related study and/or subjects who for other reasons would not be in energy balance prior to the test day, a diet standardization may play a significant role. However, when macronutrient balance is a precondition in a study it is important to standardize prior to the test day with regard to physical activity, fasting period and alcohol consumption in order to ensure equally filled glycogen stores.<sup>20</sup> So far it has not been examined whether the menstrual cycle has any influ-

**Table 3** Reproducibility of palatability scores

	ND group			D group		
	Mean difference (mm)	CR (mm)	r	Mean difference (mm)	CR (mm)	r
<i>Breakfast</i>						
Visual appeal	0	47	0.49*	-1	31	0.57***
Taste	-7*	20	0.81***	-3	35	0.44*
Smell	-6	30	0.73***	2	36	0.47**
Aftertaste	5	37	0.74***	-1	37	0.64***
Overall palatability	-4	23	0.85***	0	35	0.42*
<i>Lunch</i>						
Visual appeal	0	33	0.73***	0	28	0.67***
Taste	2	25	0.68***	0	32	0.62***
Smell	-6	30	0.53**	-4	29	0.64***
Aftertaste	-2	30	0.71***	1	32	0.76***
Overall palatability	4	20	0.79***	3	29	0.68***

ND group: subjects not diet standardized prior to test days ( $n=23$ ). D group: subjects diet standardized 2 days prior to test days ( $n=32$ ). Mean difference: mean of (test 1 - test 2). \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .

**Table 4** Power calculations for paired and unpaired design. Number of subjects needed to show the detectable difference of appetite scores. In unpaired design the number of subjects refers to subjects per group; calculations based on pooled data,  $n=55$ 

Detectable difference =	Paired design				Unpaired design	
	Power = 0.8		Power = 0.9		Power = 0.8	
	5 mm	10 mm	5 mm	10 mm	5 mm	10 mm
<i>Fasting</i>						
Satiety	70	18	96	24	88	22
Hunger	54	13	70	20	62	16
Fullness	64	16	84	21	62	16
Prospective food consumption	64	16	84	21	108	27
Salty	100	25	132	34	> 200	108
Sweet	165	42	220	56	> 200	108
Savoury	78	20	105	26	> 200	150
Fatty	100	25	132	34	> 200	115
<i>4.5 h mean</i>						
Satiety	26	< 8	34	9	88	22
Hunger	29	8	40	11	108	27
Fullness	20	< 8	27	< 8	120	31
Prospective food consumption	20	< 8	27	< 8	140	35
Salty	60	15	80	20	> 200	96
Sweet	54	13	70	18	> 200	76
Savoury	45	12	60	16	> 200	76
Fatty	37	10	50	13	> 200	96
<i>Peak/nadir</i>						
Satiety	90	22	120	30	190	50
Hunger	125	32	165	44	> 200	76
Fullness	125	32	165	44	> 200	68
Prospective food consumption	45	12	60	16	> 200	76

ence on this type of data. Until then it is recommended to test women on the same day in their menstrual cycle due to the known changes in women's energy intake, energy expenditure, and gastric emptying.<sup>21-25</sup>

Comparing CRs and correlation coefficients between identical ratings on the two different test days, it is clear that a low CR is not necessarily synonymous with a strong correlation and vice versa. For example a CR of 27 mm was reflected by an  $r$ -value of 0.16 (n.s.) in one case (fasting value of fullness in the ND group) and of 0.85 ( $P < 0.001$ ) in another (4.5 h mean of desire for something salty in the ND group). And one of the strongest correlations (0.89,  $P < 0.001$ ), between the first and second ratings of the desire to eat something savoury in the D group,

equals a CR of 29 mm, meaning that the range of the second rating in total covers about 60% of the scale. Thus, the correlation analysis should not stand alone when assessing reproducibility.

### Palatability

The meals of this study were representative of a normal Danish diet, and the design was made to assess reproducibility and validity of VAS scores. Therefore the meals were identical on the two test days and hence no differences in palatability scores were expected. However, in the no-diet group the taste of the breakfast was rated to be best on the first test day. This could be a true effect arising from being

**Table 5** Validity. Correlation coefficients (*r*) of VAS-scores (prelunch, pre–post difference, 4.5 h mean) vs subsequent lunch energy intake (EI). Both groups pooled, *n* = 55

	Energy intake		
	Day 1	Day 2	Mean
<i>Prelunch</i>			
Satiety	– 0.29*	– 0.33*	– 0.42**
Hunger	0.26(*) <sup>a</sup>	0.37**	0.32*
Fullness	– 0.34*	– 0.36**	– 0.43**
Prospective food consumption	0.36**	0.37**	0.39**
<i>Pre–post difference</i>			
Satiety	0.35**	0.35**	0.39**
Hunger	0.25(*) <sup>a</sup>	0.36**	0.33*
Fullness	0.32*	0.32*	0.40**
Prospective food consumption	0.27*	0.35**	0.35**
<i>4.5 h mean</i>			
Satiety	– 0.39**	– 0.52***	– 0.52***
Hunger	0.38**	0.49***	0.50***
Fullness	– 0.44***	– 0.51***	– 0.52***
Prospective food consumption	0.48***	0.51***	0.53***

(\*) $0.05 < P < 0.01$ ; \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ . <sup>a</sup>Outlier left out, *n* = 54. Prelunch: the rating obtained immediately before lunch. Pre–post difference: the difference of ratings immediately before and after lunch. 4.5 h mean: the mean of ratings of the 4.5 h preceding lunch.

offered an identical meal on two occasions. However, this effect was not reflected in any of the other palatability ratings, nor in subsequent appetite ratings or energy intake at lunch, which might have been expected.<sup>26</sup> Therefore, it is probably due to a type 1 error (‘false positive’ result) due to the large number of comparisons. All meals served were rated better than average.

### Power

An effect size of 10% would be a reasonable and realistic difference to look for in studies of appetite. From Table 4 it can be deduced that, with a study power of 0.8 and using a paired design, 32 subjects will be enough to test all parameters (except desire for something sweet) for fasting, mean and peak/nadir values. Further, this number will sufficiently cover a test of an effect size of 5% of 4.5 h mean values of the appetite ratings. If the effect parameters of interest are limited to fasting and mean appetite ratings, 18 subjects are enough to test the hypothesis using a power of 0.8. With a power of 0.9, 24 subjects should be included. The large variation in peak/nadir values may be explained by the fact that all subjects—despite large differences in body size—received the same amount of food for breakfast (2 MJ). This variation would probably decrease if meal size was adjusted to body size.

Considerably more subjects are needed when using an unpaired design, and the smallest number is needed is 32 subjects to discover a difference of 10 mm for hunger or fullness. This is important for interpreting studies in the literature which have failed to find any effect of experimental treatments/manipulations on hunger levels. These negative outcomes may reflect

type 2 errors arising from conducting studies with insufficient power.

### Validity

Validity is difficult to determine, since we do not have an objective measure of appetite which can be compared to the VAS scores. The validity on a short term basis can, however, be determined by correlating premeal values to subsequent energy intake. Another way of looking at validity is to make sure that a meal actually changes the value of the different VAS parameters. We therefore also correlated the change of VAS scores during the *ad libitum* lunch (pre–post difference) and meal size. In many set-ups investigation of the effect of a preload consumed minutes or hours before on a subsequent *ad libitum* test meal is desired. To address this issue we examined the relationship between the mean of the series of postprandial (after breakfast) VAS scores and subsequent *ad libitum* energy intake at lunch.

Our results showed that postprandial mean values were more strongly correlated with subsequent intake than both prelunch values and pre–post differences in appetite scores. This was not so surprising, since the mean values are less variable than just one or two single ratings (Table 1) and may relate more physiologically to the perceived desire to eat.

Furthermore, our results showed that the correlations were the same or higher on the second test day compared with the first test day. This is probably due to a training effect for subjects not priorly familiarized with the VAS scores. The most significant correlations are, however, obtained by using a mean of the two test days.

Other studies have found higher correlation coefficients for validity than in the present study. These differences can partly be explained by different ways of calculating the correlations,<sup>9,14–17</sup> which have sometimes included data from the same subject measured more than once in the analysis, meaning that within-subject variation is analysed as between-subject variation in the correlation analysis. This is usually considered to be a less than optimal statistical procedure.

On the other hand, Barkeling *et al*<sup>10</sup> tested the predictive validity of subjective motivation to eat ('desire to eat', 'hunger', 'fullness' and 'prospective consumption') on VAS and found that only the variables 'desire to eat' and 'prospective consumption' predicted forthcoming food intake. Their correlations were close to those obtained in our study.

In conclusion, despite large variations of repeatability coefficients, appetite scores done by VAS can be reproduced and therefore used in studies using single meals. However, their use requires that some consideration should be given to specific parameters being measured, sensitivity and power calculations in order to avoid type 2 errors. Neither fasting nor 4.5 h mean or peak/nadir values seem to be sensitive to diet standardization prior to the test day in this group of normal-weight men, but it might play a role in more restrained eaters. Within-subject comparisons are more sensitive and accurate than between-subject comparisons. The number of subjects needed for studies with appetite scores can therefore be reduced considerably by using paired designs. All appetite parameters seems to explain subsequent energy intake to a certain degree.

#### Acknowledgements

We are grateful to Charlotte Kostecki, Karina Graff Larsen and Berit Kristiansen for expert assistance in the preparation and serving of more than 200 meals. The study was supported by The Danish Research and Development programme for Food Technology (1995–1998) and The Danish Medical Research Council.

#### References

- 1 Blundell JE, Stubbs JR. Diet composition and the control of food intake in humans. In: Bray GA, Bouchard C, James WPT (eds). *Handbook of obesity*. Marcel Dekker: New York, 1997, pp 243–272.
- 2 Joyce CRB, Zutshi DW, Hrubes V, Mason RM. Comparison of fixed interval and visual analogue scales for rating chronic pain. *Eur J Clin Pharmacol* 1975; **8**: 415–420.
- 3 Revill SI, Robinson JO, Rosen M, Hogg IJ. The reliability of a linear analogue evaluating pain. *Anaesthesia* 1976; **31**: 1191–1198.
- 4 Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain* 1983; **17**: 45–56.

- 5 Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA. Multiple session experimental pain measurement. *Pain* 1996; **67**: 327–333.
- 6 Robinson RG, McHugh PR, Folstein MF. Measurement of appetite disturbances in psychiatric disorders. *J Psychiat Res* 1975; **12**: 59–68.
- 7 Silverstone T, Fincham. Experimental techniques for the measurement of hunger and food intake in man for use in the evaluation of anorectic drugs. In: Garattini S, Samanin R (ed). *Central mechanisms of anorectic drugs*. Raven Press: New York, 1978, pp 375–382.
- 8 Lappalainen R, Mennen L, van Weert L, Mykkänen H. Drinking water with a meal: A simple method of coping with feelings of hunger, satiety and desire to eat. *Eur J Clin Nutr* 1993; **47**: 815–819.
- 9 Porrini M, Crovetti R, Testolin G, Silva S. Evaluation of satiety sensations and food intake after different preloads. *Appetite* 1995; **25**: 17–30.
- 10 Barkeling B, Rössner S, Sjöberg A. Methodological studies on single meal food intake characteristics in normal weight and obese men and women. *Int J Obes* 1995; **19**: 284–290.
- 11 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**: 307–310.
- 12 Stratton RJ, Stubbs RJ, Hughes D, King N, Blundell JE, Elia M. Comparison of the traditional paper visual analogue scale questionnaire with an Apple Newton electronic appetite rating system (EARS) in free living subjects feeding *ad libitum*. *Eur J Clin Nutr* 1998; **52**: 737–741.
- 13 Raben A, Tagliabue A, Astrup A. The reproducibility of subjective appetite scores. *Br J Nutr* 1995; **73**: 517–530.
- 14 Silverstone JT. The measurement of hunger in relation to food intake. *Proceedings of the VII International Congress on Nutrition*. Hamburg, 1966, pp 51–53.
- 15 Hill AJ, Leathwood PD, Blundell JE. Some evidence for short-term caloric compensation in normal weight human subjects: The effects of high- and low-energy meals on hunger, food preference and food intake. *Hum Nutr* 1987; **41A**: 244–257.
- 16 Hulshof T, de Graff C, Weststrate JA. The effects of preloads varying in physical state and fat content on satiety and energy intake. *Appetite* 1993; **21**: 273–286.
- 17 De Graff C. The validity of appetite ratings. *Appetite* 1993; **21**: 156–160.
- 18 Food and Agriculture Organization/World Health Organization/United Nations University. *Energy and protein requirements*. Report of a joint FAO/WHO/UNV Expert Consultation. Technical Report Series 74. World Health Organization, 1985.
- 19 Altman DG. *Practical Statistics for Medical Research*. Chapman & Hall: London, 1991, pp 455–460.
- 20 Costill DL. Carbohydrates for exercise: Dietary demands for optimal performance. *Int J Sports Med* 1988; **9**: 1–18.
- 21 Bisdee JT, James WP, Shaw MA. Changes in energy expenditure during the menstrual cycle. *Br J Nutr* 1989; **61**: 187–199.
- 22 Lissner L, Stephens J, Levitsky DA, Rasmussen KM, Strupp BJ. Variation in energy intake during the menstrual cycle: Implications for food-intake research. *Am J Clin Nutr* 1988; **48**: 956–962.
- 23 Wald A, Van Thiel DH, Hoehstetter L, Gavalier JS, Egler KM, Verm R, Scott L, Lester R. Gastrointestinal transit: The effect of the menstrual cycle. *Gastroenterology* 1981; **80**: 1497–1500.
- 24 Notivol R, Carrio I, Cano L, Estorch M, Vilardell F. Gastric emptying and serum insulin levels after intake of glucose-polymer solutions. *Eur J Appl Physiol* 1984; **58**: 661–665.
- 25 Datz FL, Christian PE, Moore J. Gender-related differences in gastric emptying. *J Nucl Med* 1987; **28**: 1204–1207.
- 26 Hill AJ, Magson LD, Blundell JE. Hunger and palatability: Tracking ratings of subjective experiences before, during and after the consumption of preferred and less preferred food. *Appetite* 1984; **5**: 361–371.

## Appendix 1

**Table 6** Questions on appetite and desire for specific food types

I am not hungry at all	How hungry do you feel?	I have never been more hungry
I am completely empty	How satisfied do you feel?	I cannot eat another bite
Not at all full	How full do you feel?	Totally full
Nothing at all	How much do you think you can eat?	A lot
Yes, very much	Would you like to eat something sweet?	No, not at all
Yes, very much	Would you like to eat something salty?	No, not at all
Yes, very much	Would you like to eat something savoury?	No, not at all
Yes, very much	Would you like to eat something fatty?	No, not at all

## Appendix 2

**Table 7** Questions on palatability of test meals

Good	Visual appeal	Bad
Good	Smell	Bad
Good	Taste	Bad
Much	Aftertaste	None
Good	Palatability	Bad