CHAPTER 6:DISCUSSION

Chapter overview

6.1 DEVELOPMENT AND DEVELOPMENTAL EVALUATION SUB-STUDIES

From the developmental evaluation sub-studies (Chapter 3) the cognitive and other challenges involved when children complete FFQ type dietary assessments became evident. Thus the screener was once more adjusted to accommodate the findings as far as possible without jeopardizing its inherent aims and characteristics, realizing that many issues were still unresolved.

In spite of the attempt to ascertain the content and face validity of the item list by initially checking it in a consensus workshop of experts and then field-testing it in sub-study 1 (face and content validity) it was realised that the list should ideally be based on a recent survey in a group representative of the target population. This may, however, still not be a guarantee of validity: Caan et al [49] recommended this approach to improve the performance of a dietary fat screener that had, in fact, originally been developed by Block et al [44] using a data-base approach. Thus, the developmental evaluation resulted in a tool that was as well adapted as possible for comparative validation in the target group, even though neither the item list, nor the quantifications (reference portion size, frequency of intake categories and the appropriateness of the nutrient database) were claimed to be beyond debate.

Developmental evaluation sub-study 5 (food record) showed that it was feasible to integrate the task of keeping a food diary into the mathematics curriculum and it resulted in well-organised data collection in the main study.

6.2 MAIN STUDY: SAMPLE

Selection bias has been shown to be an important source of error in dietary surveys in the general population [273] and also in children.[226] In respect of the reliability study of this project (referring to measurement of internal consistency and test-retest reproducibility), this was largely ruled out, firstly, because the re-test sample was chosen randomly and based on the absence of a significant difference in final scores (in the first administration) between repeaters and non-repeaters (Table 5.3), and, secondly, because all children were included in the internal consistency part of the study with no drop-outs.

In the comparative validation part of the study the response rate for the food record was good (96%), primarily because of the mathematics (school) context in which it took place. This is in contrast to many studies where poor compliance and a suspicion of non-response bias have been raised with the use of food records. The act of recording could, however, have altered the

children's eating habits, a known limitation of food records. For the second reference method (parental completion of screener) the response rate was 72%, which was also considered reasonable. Overall the study was characterised by very few missing values, in contrast to some previous studies involving school children (for example reference [265]), but the external validity was limited by the fact that only one school was included in this stage of the project. Restricting the research to one grade level taught by the same mathematics teacher avoided potentially confounding (intellectual) developmental and administration factors.

Anthropometrically, on average, the participants exceeded median reference indices (weight, height and BMI for age) using the CDC 2000 growth charts as basis. The international cut-off points corresponding to 25kg/m^2 at age 18 for 12.5 year olds are 21.6kg/m^2 and 22.1kg/m^2 for boys and girls respectively.[274] Based on this, the participants' mean BMI's of 20.4kg/m^2 and 20.6kg/m^2 for males and females respectively, were interpreted as being in the 'healthy' range. Thus, in general, the sample could be taken as being anthropmetrically reflective of a population of healthy children.

6.3 TEST METHOD

6.3.1 Internal consistency

Internal reliability, also called homogeneity or uni-dimensionality, reflects the extent to which individual items in a test measure similar characteristics.[231] Therefore it has been reasoned that variance among scores in an internally reliable instrument indicates subject differences and not error.[245]

Keller et al [184] suggested that items with a corrected item-total correlation <0.2 are less relevant for measuring the construct of interest. Since the lowest item-total correlation in this study was 0.35 (Table 5.2), it was concluded that none of the food categories needed to be discarded or rephrased. Cronbach's coefficient alpha values below 0.7 indicate an excess of nuisance items or too few items in a scale. Values >0.7 indicate there are items measuring essentially the same thing and that some are unnecessary.[184] Again it was concluded that all original MEDFICTS items could be retained and that they fully covered the construct under investigation, namely fat intake. The alpha obtained in this study was higher than those included in the review by Yaroch et al.[200]

Several researchers (for example references [233, 236, 246]) have used item total correlations as an indication of convergent validity (a form of construct validity) in nutrition research. In line with

this reasoning, it could be concluded that the test method, by being internally consistent, also exhibited convergent validity.

Considering that scale consistency in classical test theory is a function of the number of items in the scale,[231] the various coefficients of internal consistency obtained for this ten-item tool are encouraging for the test method as a screener.

6.3.2 Test-retest reproducibility

The second form of reliability investigated in this study was repeatability or stability over time. If the tool is used with its original aim in mind (that is to assign individuals to the three dietary 'Steps' of the American Heart Association [203, 205]) then reproducibility of this classification is important. In this study the step classification was reduced to 'high fat' and 'prudent' as children in the general population were targeted. In over 90% of cases the first and second administration resulted in an identical classification. It was thus concluded that the screener exhibited classification reproducibility, which would be the prime concern in clinical decision-making.

If, on the other hand, the screener is used for comparing, ranking or monitoring groups of individuals, then reproducibility of the final scores might be a more useful basis for describing reproducibility. Positive correlation coefficients ('reproducibility correlations') between two administrations of dietary assessment tools have often been used as indicative of test-retest reproducibility.[275] Following this line of reasoning, one could conclude that the dietary fat screener exhibited limited (r=0.36), yet statistically significant (P=0.02) reproducibility in grade six learners as a whole.

The correlation coefficient obtained in this study is substantially lower than those reported for most nutrients by Anderson et al [276] who used a comprehensive quantitative FFQ questionnaire in Norwegian adolescents with mean age 17 years, and for the foods reported by Metcalf et al [277], who had a very heterogeneous group (one to 14 years from different ethnic backgrounds) with a mix of parent, caregiver or child as data source and a reference intake period of four weeks. Also, Andersen et al [251] found (among Norwegian 6th graders) consistent responses (r=0.62 to 0.83) on a 24-h recall and a FFQ both re-administered 14 days apart when the previous day's fruit and vegetable intake had to be recorded. By contrast, Yaroch et al [278] administered a picture sort FFQ questionnaire twice during a two-week period to low-income, overweight, African American adolescents, and reported correlation coefficients ranging from 0.28 to 0.36. The

reproducibility coefficient of the Youth and Adolescent Questionnaire ranged from 0.26 to 0.58 for different nutrients, and from 0.39 to 0.57 for foods.[279]

Thus, based on reproducibility correlations between final scores, the findings in this study are similar to some and different to other test-retest studies in children, but the obvious differences in research contexts, designs and outcome measures used in the analyses must be kept in mind. The range of correlations for FFQ's in school children reviewed by McPherson [25] was from -0.06 to 0.91, complicating generalisations. Thus, the warnings by Altman [280] and Bellach [275] that using a correlation coefficient to measure agreement may be a "misconceived" or "archaic" analysis, seem appropriate.

Looking at several statistics and exploring the distribution and structure of the measurement error has consequently often been recommended.[111, 154, 275]

The finding that the reproducibility coefficient was non-significant (P=0.29) for boys, whilst it was highly significant (P=0.008) for girls, illustrates that the measurement error may not have been equally distributed amongst participants.

The final score in the screener was the sum of the ten category scores, which, in turn, were the product of the scored portion size and frequency of intake estimations. Thus, focusing on the test-retest reproducibility of the latter two, would further explore the error structure and explain the final score reproducibility. As shown above (Table 5.4), for portion size and weekly intake the percentage agreement was over 50% in eight of the ten food categories. However, when corrected for chance (kappa statistic) the agreement was poor for almost half of the food categories, but overall, the non-agreement tended to be symmetrical. This means that about equal proportions of children changed from a higher to a lower estimate and vice versa.

The kappa's in this study are lower than those reported by Smith et al [255] where middle school students completed a 40-item checklist of foods high in total fat, saturated fat and sodium twice on the same day. These researchers reported kappa values ranging from 0.66 to 0.93. The short period between their administrations may explain the discrepancy. Jonsson et al [267] reported amongst school children 'good' percentage agreement (between 58% and 86%) of usual choice of four different breakfast foods on two occasions eight weeks apart. Based on different statistical analyses, they explain the cases of unreliability in terms of a combination of random

and systematic error. As will be evident from the following discussions, this may also be the case in this study.

It is tempting to deduce that methods agree because they are not significantly different. This approach to establishing test-retest reliability has been used in a number of reproducibility studies. Examples include Burden et al [281] (even though they did not report the statistical significance of their findings) and Anderson et al, [276] who found that the first measurement gave significantly higher values than the second. Cullen et al [282] as well as Buzzard et al [265] also reported higher mean consumption estimates in the first administration of their respective screeners. According to McPherson et al [25] this appears to be a trend for FFQ type dietary assessment in school children. Thus in this respect the present study's findings differ from previous reports. The mean difference between the first and second administration for girls was higher than for the group as a whole (2.95±25.42 compared to 0.69±32.59; Table 5.5), but also this value did not significantly differ from zero.

Whilst the small mean difference in final scores (0.69 see Table 5.5) indicates that the two administrations of the dietary fat screener in this study agreed well *on average*, the measures of variability (for example the standard deviation of the difference of 32.6 and the 95% confidence interval ranging from -9.9 to 11.3 for the difference in final score) suggest that for *an individual* absolute agreement of the final scores was less likely. This is confirmed by the Bland-Altman plot (Figure 5.1b) and non-agreement was particularly true for boys. In the case of girls there tended to be less variability (less random error), but a bias (systematic error) towards lower scores in the second administration.

6.3.2.1 Factors affecting reproducibility

Respondent and methodological factors can affect reproducibility.[80, 154, 220] Both of these could have been at work in this study.

6.3.2.1.1 Respondent factors

In the case of children, their cognitive abilities[90], specifically to record, remember or generalise their intake [29] as well as their restricted knowledge of food and food preparation [30] and limited motivation and attention span [81, 283] are well-documented child-specific respondent factors, which can contribute to error. However, true inter- and intra-individual variability also affect the measured reproducibility, for example, in five to 17 year old children the ratio of intra:inter subject variances in intake is, in general, approximately twice that observed in adults [81] and

specifically fat intake and fat practices have been shown to vary by meal and day in grade four to six school children.[284] Gender has sometimes emerged as a differentiating respondent factor, [277, 279, 285] but results are conflicting. Age, obesity and weight consciousness are additional respondent factors related to valid dietary assessment.[81, 283, 286] This cannot necessarily be traced to lack of reproducibility, since data may be reproducible, yet invalid. For example, Frank et al [285] found amongst twelve to17 year olds that age did not influence test-retest agreement, yet Bandini et al [287] found that from age ten to15 years girls tended to report energy intake less accurately. The group of children in the present study was very homogeneous in terms of age, educational level and culture. The check whether BMI was related to differences in final scores in the current data set revealed no significant correlation between BMI for age Z-score and difference in final scores in the two administrations, neither for the group as a whole, nor for the genders separately. Thus the fact that in this study the boys in the re-test sample had a BMI for age Z-score of 0.64 (compared to 0.25 for girls, see Table 5.1) did not explain the poorer reproducibility in the boys as sub-group.

6.3.2.1.2 Methodological factors

Amongst the methodological factors in a reproducibility study, the period between administrations remains controversial. Whilst too short intervals will result in learning, carry-over, or recall effects, true changes may occur if the period is too long. Frank et al [285] included a two-hour and two-week interim period in their test-retest reproducibility study in children and found lower agreement in the two-week period. They concluded that a two-week repeat measure tests variability within an individual's eating pattern, rather than the reliability of the instrument. However, most researchers recommend periods from four to eight weeks when assessment of usual diet is the aim.[80] Thus, the higher correlations typically found with briefer intervals,[25] should be interpreted with this in mind. The fact that between the two administrations in this study food recording was done for the validation study, could have affected the second administration. The food categories included in this screener were not so much prone to seasonal variation as fruits and vegetables, which have been highlighted by Joachim [288] as important factors affecting reproducibility.

The type and design of a dietary assessment tool affects reproducibility. Hoelscher et al [289] and Buzzard et al [265] found that composite food items (for example consisting of numerous foods) had poorer reproducibility than single food items. Since all items in this study were composites, this could have affected reproducibility coefficients. The scoring system is another methodological aspect that could have affected reliability. The fact that the screener permitted

quite a bit of variability (because portion size had to be specified - in contrast to qualitative and semi-quantitative FFQ - and the scoring system as such, particularly in respect of weekly intake and the multiplication principle) could have reduced reproducibility of the final scores. On the other hand, the classification of final scores into the two classes of fat intake might have increased test-retest agreement. The testing, standardised data-collection and quality control during coding minimise these two factors as strong contributors to apparent poor test-retest reproducibility in terms of final scores in this study. The statistical techniques used and the outcome variables, on which results are based, are analytical factors affecting conclusions in reliability studies, as also evident from the findings of this study.

Reproducibility is a function of sample, geographical and time factors [221], thus again limiting the generalisability of the findings of this study. In addition, precision of differences observed (Table 5.7) is also influenced by sample size. Thus, the relatively small sample size (n=39) also explains some of the variability and consequently the limited reproducibility.

6.3.3 Reflection

Some of the observations made during the administration of the screener are summarised in Table 6.1 in terms of perceived strengths and challenges.

For some food categories, the food list of the test method relied on the ability of the children to differentiate between high fat and low fat versions of outwardly similar foods, for example, for milk and table fats (different types of [tub] margarine). It may well have been that the children were unsure about the type usually eaten, even though the pictures proved very helpful. Thompson et al [118] have recommended that, when different forms of a food exist, it might be helpful to first ask about consumption of the whole food, for example milk, and then the proportion of times each form is consumed (that is 'nesting'). In the present study the filter question "Do you eat foods such as those on the picture?" was intended to fulfill a similar function.

The grouping of items has previously been identified as a FFQ design issue in the sense that multiple, separate questions appear to result in greater accuracy.[118] In screening the aim is rapid assessment, which essentially means losing detail. Ideally this should not be at the expense of accuracy, but in reality it would mean striking a balance. On the other hand, earlier publications have reported that lengthy FFQ's may overestimate intake.[268]

Initially (typically for the first two food categories) the children needed considerable time and very clear, repeated instructions to code their responses on the answer sheets. Then the process speeded up. It was considered to add pictures to the answer sheet, but following consultation with the teacher this was not done, because then learners could potentially work at their own pace, without being briefed about the distinguishing features of the category, thus affecting validity and reliability, and disrupting orderliness.

TABLE 6.1: STRENGTHS AND CHALLENGES OF PRACTICALITIES OF SCREENER ADMINISTRATION

	Strengths	**Challenges**
Venue	• First impression of a 'special' occasion: The conference room was not children's usual domain and evoked comments (for example the comfortable chairs)	• Artificial; not realistic
Group size	• Even though primarily determined by practical constraints (that is size of conference room), group size proved to be ideal from the research perspective	• Each class was divided into three groups, with implications for responsible time management
Setting within venue (U-shape with separators)	• Interviewer could unobtrusively check coding, ensure sustained participation, keep eye-contact, and reduce omissions • Not school-like • Privacy appreciated by children	• Repeatability in other school contexts
Method of administration (Interviewer-guided and-paced; demonstration of coding)	• Data quality • Orderly approach appreciated by school administrators • Can stick to available time (one school period per administration) • Consistency • The visit by a 'dietitian from the university' added status to the children's involvement	• Different work pace of different children (see below discussion)
Administration in school time	• High participation rate	• Can be perceived as disruptive, unless meaningfully integrated into school programme • Requires commitment from involved teachers
Appearance of tool	• Colour and real life pictures were appreciated (visual appeal [290]) • Concrete and realistic • Spontaneous reaction to the pictures. This included positive and negative comments (for example for snacks and organ meat respectively)	• Not all brands pictured • Brands change • Cost • Comments take time and can result in peer influence

Among the practical problems observed with the screener as a FFQ type assessment tool are the following:

- In spite of the example given, facial expressions sometimes suggested that the cognitive challenges associated with reporting frequency of intake were real.
- The children needed repeated reassurance about how to code very low consumption frequencies. This observation may partially explain the general finding that FFQ's tend to overestimate intake.[25]
- Some children found it difficult to separate frequency of intake from a particular portion size, for example milk consumed as drink, on cereals and in coffee / tea. This appears to be different from Thompson et al [118] who claimed that, in general, asking about frequency of intake and about portion size versus frequency of intake of a 'standard' portion size made no difference.
- Linking the word 'medium' to the reference portion size might have implied 'average' or 'normal' or 'recommended'to some children.

In general, many of the cognitive challenges involved in responding to FFQ's as described by Subar et al [117] for adults, were observed during the administration of the test method, in this case to children.

Raat et al [291] used response rate and missing answers as indicators of feasibility, when they tested and validated the Child Health Questionnaire. If these criteria are applied to the current study, it can be concluded that the feasibility was high since the response rate was very high with almost no missing values. The physical setting (that is school context and venue) and the data collection approach (small group, structured interview and coding with teaching aids) probably greatly contributed to this outcome.

All data collection regarding the dietary fat screener had been done by the researcher, a registered dietitian, personally. The characteristics of screening stipulate that a screening tool should be administrable by any qualified professional (that is individuals who are qualified by virtue of their education, experience, competence, or privileges). Thus, whilst the approach ensured consistent administration, it limits conclusions regarding generalisability to other and different health care workers. Inter-rater reliability consequently remains to be established. The standardised text should, however, be helpful in this regard.

The reference period for the test method was since the beginning of the year (that is "since you were in grade six"). This was reflective of about nine months of the year and was assumed to reflect usual intake, since fat intakes are not so much prone to seasonal variations. Furthermore, it was assumed that for South African school children this clear and explicit time frame would be meaningful as it coincided with a cognitive reality, that is the school year. It is, nevertheless, realised that Wolfe et al [239] found that such a relatively long time frame may cause problems for participants.

The fact that trouble was taken to highlight distinguishing feature(s) of each food category helped to avoid interpretation problems and ensure that the grouping of items was clear to the participants, as recommended by Wolfe et al [239] and Livingstone and Robson.[81]

6.4 REFERENCE METHODS

6.4.1 Food record

6.4.1.1 Plausibility of energy intake data

Since the three-day food record was chosen as primary reference method and because it was the only quantitative reference method in this study, it was considered important to establish its plausibility. Reported energy intake has often been used as surrogate indicator of the total quantity of food intake.

In contrast to micronutrients, there are no biochemical markers of energy intake. Three methods of validation are currently available in respect of energy intake data: [159]

- Comparison of self-reported energy intake with the energy intake required to maintain weight.
- Direct comparison of reported energy intake and measured energy expenditure (for example the doubly labeled water technique).
- Comparison of reported energy intake with presumed energy requirements, both expressed as multiples of basal metabolic rate.

Technical and cost considerations excluded the first two methods for use in this study. Consequently the latter method, developed by Goldberg et al,[46] was employed. This consists of the so-called cut-off 1, which tests whether reported energy intake can be representative of long-term habitual intake, and cut-off 2, which may establish whether reported energy intake is a plausible measure of the actual diet during the measurement period. In the Goldberg method the reported energy intake is judged against presumed energy requirements by expressing the energy

intake as a multiple of the estimated basic metabolic rate (BMR). The ratio is referred to as the physical activity level (PAL).[46, 159]

Although the concept was originally developed for adults, the principle has been applied in studies with children, for example Torun et al [292] and O'Connor et al.[293] A PAL of 1.06 has been used as cut-off to test whether reported energy intake from a three-day record is a plausible measure of the food consumed during the actual measurement of dietary intake. [43, 293] Nevertheless, Livingstone et al [294] recently cautioned against its application for identifying individual misreporters in paediatric groups.

Consequently the PAL was used in this study to establish whether overall bias at group level was present and not to discard the twelve individual records below 1.06. The following additional, inter-related reasons are presented for this decision:

- It has been claimed that excluding underreporters may introduce an unknown bias into the data set,[159] because underreporting is the result of undereating (eating less whilst recording) plus underrecording (failure to record everything that was eaten).[295]
- Omissions and intrusions have been found to result in low accuracy and low consistency in grade four children's dietary recalls,[296] but misreporting (under- and/or overreporting) may also be selective in terms of certain nutrients or foods.[56, 243]
- Underreporting appears to be not equally distributed within a population. Weight status, sociodemographic and psycho-behavioural factors may all be predictors of underreporting [159, 230] and contribute to the so-called subject-specific bias which seems to be characteristic of specific individuals regardless of the dietary assessment method used and persistent over time,[297] as discussed in the review of literature (Figure 2.3).
- Finally, after critically assessing the sensitivity and specificity of the original Goldberg cut-off, Black [298] concluded that information of each subject's activity level is also necessary to identify diet reports of poor validity.

The mean PAL of 1.45±0.39 found in this study thus suggested that, on average, the energy intake was plausible for the reporting period. The value, however, was lower than those related to light habitual physical activity, namely 1.54 and 1.48 for six to 13 year old boys and girls respectively.[292] A PAL >=1.4 but <1.6 for boys and girls between ages nine and 18 (within the 5th to 85th percentile for BMI) corresponds to 'low active' in the four physical levels published by the Institute of Medicine.[272]

A bias in dietary reporting is thus probable also in study, similar to previous studies in children [25, 81, 159, 299, 300] where often doubly labeled water acted as reference. Thus, the energy intake could not be regarded as representing long-term habitual diets.

Underreporting may be intentional and unintentional. In general, food recording can be subject to problems like changes in intake to simplify recording, decisions not to eat or record foods that might be considered 'undesirable' and failure to record food that were eaten in excessive amounts.[243]

Factors that could, in addition, have contributed to the underreporting specifically in this study include the following:

- A tendency toward recording fatigue, evident in consecutive day recording, could be noticed as, overall, mean energy intake on the third day was significantly lower than on the first day of recording (Table 5.9).
- Weekend days were not proportionally represented and it was found that energy intakes were always higher on weekend days compared to the other days of recording of a particular recording group, regardless of whether the weekend day was the first or the last day of recording. For the two recording periods that included a weekend day no recording fatigue was shown. This was surprising for the Sunday to Tuesday group. Thus it could also be that during weekends parents became involved and affected the recording 'pattern'.
- Whilst the use of the electronic scales was shown during the developmental evaluation sub-study to be associated with more comprehensive *descriptions* of food consumed, the results in the main study indicate that the inconvenience of weighing (in contrast to estimating) may have resulted in underrecording and / or undereating, because the children weighing their food recorded lower intakes than those using household measures. Almost double the number of children weighed their foods, thus affecting the mean energy intakes.
- The fact that for girls a non-significant correlation between weight and mean energy intakes was found, confirmed the presence of selective (gender-specific) personal bias,[292] since higher habitual energy intakes can be assumed for persons with higher weights. Based on the PAL, the percentage of DRI for energy and the correlations between body weight and reported energy intakes, it thus appeared that the boys' food records were more plausible than the those of the girls.

- The BMI z-score of the girls in this sample (0.43±1.1) suggests that these girls' BMI's were, on average, higher than the reference population, also a variable known to be associated with underreporting.[297, 299]

It follows that a coalescence of methodological (that is food record-related) and respondent (personal) factors seemed to have played a role in the underreporting. This is in line with previous studies on the characterisation of low energy reporting for example Cook et al [301] and as indicated in Figure 2.3.

The observed day-to-day variability in intakes of energy, fat, saturated fat and cholesterol relative to international standards, that is standard deviations and coefficients of variability from then Continuing Survey of Food Intakes by Individuals,[272] suggested overall comparability of this study to other publications.

6.4.1.2 Reflection

Whilst the strengths of the food record (see review of literature) explain why it is generally used as reference method, the drawbacks limit its use. In this study an attempt was made to administer the food record in such a way to minimise some of the weaknesses without losing the strengths. The following was noted in respect of the approach and context used:

Firstly, the respondent burden associated with completing food records is well known. By integrating the food recording into mathematics, the process became a school assignment and not a task to be completed over and above the daily chores of learner and teacher. Nevertheless, commitment from the responsible teacher remained crucial. Helping the teacher set up meaningful follow-up assignments was a form of recognition and an attempt to contribute to curriculum development. The current paradigm shift within the South African education system could thus be used to create a win-win situation for nutrition research and the involved school. It is realised that the context as a whole may have created some pressure for the individual learner to participate.

Secondly, the fact that all learners were included addressed selection bias, which is very often associated with food records, because those who complete the records may be significantly different from those who do not participate.

Thirdly, since the whole class was involved, the act of recording did not make the participant feel 'out' when weighing and recording foods. It remained, however, important to maintain the motivation and ensure that (negative) peer pressure resulted in not (properly) doing the task at hand, overdoing it or in terms of changing usual eating habits. Again the daily reinforcement and guidance by the teacher was critical; also for checking that weighing and recording techniques were adequately performed.

Providing scales and household measures added value to the mathematics learning experience, in addition to stimulating the children. The logistics, detailed learner training, briefing and daily re-briefing were perceived as time consuming, particularly when the food recording extended beyond one day and the learning benefit related to mathematics decreased.

Employing a dietitian of the same culture as the children, who was up to date with their eating habits, language usage and trends in the food industry, proved to be very valuable, since the handwritten records showed great variations in the participants' language usage (terminology used and spelling abilities, for example “kaaikiebotter” [peanut butter], “myjenys” [mayonnaise], “niekerball” [sweet], “stuee” [stew)]) and participants often recorded foods by their brand names.

Including measures of quality control (in terms of anthropometric and design factors) proved valuable for explaining results. Future investigators could consider refinement by combining estimated BMR with physical activity measured objectively (for example with a triaxial accelerometer [302]).

In general, whilst for boys the test-retest reproducibility of the test method was shown to be worse than for girls (see previous discussion), the credibility of the food records was better for boys than for girls. This might be explained in terms of gender-related food awareness: In the test method, which relied on memory, the presence of random error suggested that the boys were guessing, whilst the girls tended to be more consistent (but perhaps consistently underreporting). When the task at hand was to record intake, the boys appeared to be closer to the ‘truth’ and the girls still underrecorded. This implied a gender-specific error structure: For boys poor comparative validity of the test method should thus already at this stage be expected, because of the low reproducibility of the test method, whilst for girls the potential of agreement with the reference method existed. However, there was a good chance that for girls both, the test and the

reference method, reflected underreporting, and both would thus be 'untrue': The dilemma of a 'non-golden' reference method.

6.4.2 Screener by parents

Reporting on fat intake requires some knowledge of foods, food preparation and food composition (in terms of brands and food labels) in order to differentiate between, for example, different types of meat, dairy products, table fats and cooking methods, all of which children might not have. Therefore, using parents as surrogate sources of information in the comparative validation was deemed important.

The findings of the comparison between children and parents are discussed in detail as part of the following sections.

6.5 COMPARATIVE VALIDATION

6.5.1 Basic associations between test method and reference methods

The processing involved in the test and reference methods allowed for several approaches and outcome variables to be used in the comparative validation. In the following section the results of the comparison of the test method to the reference methods in respect of the non-classified final outcome are discussed and interpreted. Since no dietary assessment method is perfect, a critical analysis and understanding of the nature and source(s) of the underlying error forms the backbone of comparative validation.

6.5.1.1 Test method versus food record

Against the background of limited reproducibility of the test method amongst boys in this study, the lack of statistically significant correlations between the final score obtained in the screener and the three outcome measures from the three-day food record (PFE, PSFE and cholesterol) was not surprising. Equally, the relatively low (yet statistically significant) test-retest reproducibility coefficient for the group as whole (based on final scores in the test-retest assessment), partly explains the absence of an association between test method and the food record. Thus, since reproducibility is a requisite for validity, only for the girls a meaningful comparative validity could be expected.

The aim of MEDFICTS is to predict high PFE, PSFE and cholesterol intakes. The current study did not yield significant correlations between the final score of the test method and any of these measures of fat intake for the group as a whole. This is in contrast to the results of Srinath et al,

[201] Kris-Etherton et al [203] and Taylor et al.[205] Possible reasons could be the nature of the current target group (age and the effect thereof on both, the test method and the reference method, as well as cultural differences) and/or the modifications made to the tool.

The finding that total fat intakes (rather than energy contributions) revealed significant correlations, has previously been documented.[199, 206, 217] Van Assema et al [199] explained it by very high correlations between total fat intake and total saturated fat intake resulting in a low variance in percentage energy from fat, which could lead to low correlations between test method final scores and measures of fat intake from the food record. A further explanation is probably the fact that the test method did not include non-fat energy sources.

Rohrmann and Klein [264] also used total intakes (as opposed to energy contributions) as outcome measures in their validation study of a dietary screener. They reported correlation coefficients of 0.44, 0.50 and 0.56 ($P<0.001$) for total fat, saturated fatty acids and cholesterol respectively. The items in their short questionnaire were based on a representative food consumption survey of the target population, they did not measure portion size, the target group consisted of adults, and the reference method was a 148-item FFQ, all in contrast to the present study and all of which could offer an explanation for the higher correlation coefficients they obtained.

Caan et al [49] evaluated the performance of a dietary fat screener and reported that it was more effective at classifying respondents into quintiles of total fat intake than into quintiles of percentage of energy from fat.

The limitations of using correlation coefficients to establish the validity of a dietary assessment method have been mentioned before and are well documented.[27, 88, 154, 270] The main problem is that it cannot be judged on a null hypothesis basis. Furthermore the confounding effect of intra-subject variation on usual intakes is not taken into account and consequently, because the reference method in dietary assessment itself is usually imperfect, a correlation coefficient may underestimate the level of agreement (attenuation bias) with the actual usual intake. The reason for reporting correlation coefficients in this study is thus primarily for comparing results to previous research. Furthermore, in a study comparing different statistical methods for assessing relative validity of a FFQ, Spearman correlations were found to be useful [303] and the application of a combination of statistical methods was again highlighted.[304]

PFE and PSFE intakes of children classified by the test method as consuming a high fat diet were very similar to those classified as prudent eaters, but absolute intakes (energy, total and saturated fat, cholesterol) were higher in the high fat group. This suggests lack of concurrent validity,[198] when two of the intended outcomes of the test method, that is PFE and PSFE, were used as standard. As in the case with the correlations discussed above, the test method tended to be stronger related to absolute intakes compared to proportional intakes.

6.5.1.2 Test method versus screener by parents

As indicated in Chapter 5, when the screener was in the hand of the parents in respect of their grade six child, it also exhibited characteristics of homogeneity. The findings from the item total correlations, Cronbach's alpha and the split half method were comparable to what was found when the screener was used by the children themselves to perform a self-assessment of intake (see previous section on test method). It is thus concluded that the dietary fat screener *per se*, regardless of the data source (that is the children as primary informants or the parents as surrogates) was internally reliable, which could be taken as enhancing its content validity.[231] In the original pilot testing of MEDFICTS Srinath et al [202] had already noted that self-administered and interviewer-administered application of the tool resulted in similar findings.

If, on the other hand, the outcomes of the measurement by parents of their children's diets were compared to the outcomes obtained from the children themselves, the following was found: Firstly, when the two primary building blocks of the screener (that is the reported usual portion size and weekly consumption) were analyzed (Table 5.16), it appeared that:

- In general (for both, portion size and weekly consumption, as well as across the food categories) there was limited agreement between parents and children. This was reflected by relatively low percentages of identical responses and also few food categories with fair or moderate chance corrected agreement. In three cases the kappa value was in fact negative, meaning worse than chance agreement. Hoehler [305] has argued that the presence of bias reduces kappa values. As evident from the following discussion there appeared to be systematic error in terms of reported portion size and weekly consumption, thus explaining some of the very poor chance corrected agreement obtained between children and parents.

- There appeared to be a systematic error in the sense that most of the non-agreeing responses were not symmetrical (McNemar data in Table 5.16). Typically parents reported smaller portion sizes and less frequent consumption. This seemed to have had a

'carry-over-effect' to the category and final scores as evident from Table 5.17 and Figures 5.3. The reason for the lower scores by the parents may be either due to underreporting by the parents or overreporting by the children, or both. For preschool children it has been found that mothers were more likely to underreport then over-report foods.[81] In this study the difference appeared to be more evident for the food categories to which the parents may have had an 'unhealthy' connotation (for example dessert, convenience foods, fried and baked foods, as compared to meat, milk and cheese) whilst being popular (and possibly overreported) by the children as suggested by Koehler et al. [92] This does, however, not explain the finding in respect of snacks.

- Whilst reported frequency of intake *and* portion size showed lack of agreement between parents and children, this appeared to be more evident in respect of portion size. The lack of symmetry was equally common in reported portion size and weekly consumption.

- Thus, at least one of the data sources (children or parents) appeared to lack validity in terms of reported portion size and weekly consumption using the dietary fat screener. Matheson et al [306] assessed the validity of eight to twelve year old African American girls' self-report of food portion estimates and found "sizable errors in quantitative estimates". Consequently cautious interpretation of the children's self-reports of portion estimates seems to be necessary. On the other hand, from the food records it was evident that the children made many food choices in the absence of their parents, making parental error also not unlikely.[81, 307]

Secondly, the correlations between parents' and children's individual category scores (which are the product of the scored portion size and weekly consumption) were statistically significant for only three, borderline for two, and non-significant for five of the ten food categories.

Thirdly, the correlation between the sums of the category scores (that is the final scores of parents versus children) was small (r=0.23), yet statistically significant (P=0.04). A differential pattern emerged for boys and girls, with parents and their sons not showing a linear relationship. Since the boys' final scores had also not been reproducible (see previous discussion) this is not surprising, as reproducibility is a prerequisite for validity. In the case of girls, the statistically significant positive correlation coefficient (r=0.33, P=0.04) between the parents and their daughters only shows that higher values in the one group were associated with higher values in the other group.

The finding that the mean difference between parent-daughter pairs' final scores differed significantly from zero confirmed earlier suggestions (see frequency of intake and portion size) of systematic error. This was also evident from the plots of the individual data points (Figures 5.3a and 5.3b), which additionally show the wide scatter around the diagonal (Figure 5.3a) or horizontal zero line (Figure 5.3b) of perfect agreement, suggesting considerable variability for individual pairs.

In spite of a relatively high percentage (76%) of identical classifications into high fat or prudent intake, the chance-corrected agreement between parents'and children's final classifications was poor. According to Hoehler [305] this may have two reasons: Firstly the presence of bias, which was shown to be the case in this study and, secondly, by prevalence effects, which also applied to this study. The latter occurs when the model is based on an underlying continuous variable; in this case the final score.

6.5.2 Classification agreement

In many situations the aim of the test method as screener would be primarily to classify individual intakes into 'high fat' versus 'prudent'. In this case classification agreement between the test method and the reference methods would be of prime interest.

A tool's overall predictive value is defined as its ability to predict correctly the presence or absence of nutritional risk. Consequently the dietary fat screener's overall predictive value refers to its ability to predict correctly the presence or absence of high fat intake (based on references [45, 308]). The percentage perfect classification agreement between the test method on the one hand, and the two reference methods on the other hand, provides an overview of the test method's overall predictive value and thus an indication of criterion-related validity. Figures 5.4a to 5.4e graphically represent the findings from the triangulation and are discussed below (based on n=72), even though the classification agreements between the test method and each of the reference methods separately (based on n=93 and n=78 respectively for reference method 1 and 2) were also presented in the results section.

The screener, when completed by grade six children or their parents, showed some classification agreement (about 74%). Percentage agreement between the test method and food record tended to be similar or slightly higher than this, when PFE, PSFE or any of the three measures of high fat intake acted as criterion (71, 78 and 85% respectively). When cholesterol intake was included

as reference, the agreement was much lower (about 20%). The agreement between the two reference methods followed a similar pattern, but was always less than the corresponding agreement between the test method and the food record.

Whilst the three-day food record in this study was shown to be plausible in respect of the recording period, the PAL values obtained suggest that it was probably not reflective of 'usual' energy intake. Consequently it may also have been inadequate in capturing 'usual' cholesterol intake. Furthermore, cholesterol (compared to total fat and saturated fatty acid) intake is known to have a high intra-subject variance for the general population [27] and even more so for children (five to 17 years old).[81] The number of days required, on average, to estimate true usual adult intakes of cholesterol might be as high as 139-200 and 13-15 days for an individual and a group of individuals respectively. [167] The test method and the screener completed by the parents showed poor agreement with the food record classification based on cholesterol intake, yet they tended to agree (in terms of percent identical classifications) with one another. Thus, it is concluded that the three-day food record may have been an inappropriate reference method in this respect, rather than the test method being an inadequate tool for screening for high usual cholesterol intakes.

6.5.2.1 Sensitivity and specificity

Measuring sensitivity and specificity to describe the validity of dichotomous screening tests is very common in the medical literature. [309, 310, 311] It is also increasingly used in nutrition research (for example references [33, 49, 193, 312]).

Even though no dietary assessment method is flawless, the weighed food record is an accepted, practical relative standard.[157] Consequently the criterion-related validity of the test method in terms of the sensitivity and specificity was expressed relative to the various measures of fat intake obtained from the food record (see Table 5.14).

Sensitivity is the ability of the test method to correctly identify individuals truly at nutritional risk, that is true positives. Thus, in the present context it would refer to the dietary fat screener's ability to identify correctly children who, according to the three-day weighed food record, had a high fat intake.

Specificity measures the test method's ability to correctly identify persons who are not at nutritional risk, that is true negatives. In analogy it would refer to the dietary fat screener's

ability to identify correctly the children who, based on the three-day food record, consumed a prudent diet.

From Table 5.14 it is evident that for the dietary fat screener with a cut-off value of 68 there would be considerable misclassification of those not at nutritional risk as determined by the food records and to a much lesser extent of those at nutritional risk, because the dietary fat screener exhibited high sensitivity in identifying high fat intakes, but lacked specificity. In the clinical medicine context, a large sensitivity means that a negative test can rule out the disease (David Sacket coined the acronym “SnNOut” for this); thus, for the dietary fat screener with its high sensitivity it could be concluded that a child for whom the result indicated prudent intake, high fat intakes could be ruled out. Equally, a large specificity would have meant that a positive test could rule in high fat intakes (The David Sacket acronym for this: "SpPIn"). The latter was, however, not the case for the dietary fat screener in the current study.

The ideal would be to have both, high sensitivity and high specificity, in a screening tool. It is, nevertheless, well known that in real life usually a balance must be struck between sensitivity and specificity, specifically when the test variable is a continuous variable. Decisions on the appropriate cut-off for a screening test mainly depends on the consequences of identifying false negatives and false positives,[45] but also on the implications of the test for the patient and the health care system and availability of effective treatment.[310] The purpose of a particular study may also play a role, for example if the test method is used to establish prevalence the aim may be to have a balance between false positives and false negatives. In a study that compares rates in different populations the absolute rates may be less important, but the primary concern would be to avoid systematic bias, meaning a specific test may be preferred, even at the price of some loss of sensitivity.

In the context of this study, clearly it would be undesirable to have many false negatives, that is, failure to identify those who are really at risk of high fat intakes, because the benefits of early detection and intervention that are associated with preventing CNCD in childhood [15, 19, 20] would be missed. It has been argued that general screening tools aimed at detecting malnutrition should primarily be sensitive, because an in-depth, follow-up assessment to confirm the screening result is presumed, or because it is reasoned that giving nutritional care to those who do not need it does no harm. Furthermore, early diagnosis generally has intuitive appeal, as earlier treatment is thought to be related to improved prognosis.[310]

The practical and cost implications of unnecessary screening, follow-up and interventions should, however, also be considered, apart from potential harms related to anxiety, adverse effects of labeling (stigmatisation and discrimination), inconvenience and the possibility of childhood risk factors not tracking into adulthood (inconsequential risk).[309, 310, 313] From the above it follows that a high proportion of false positives is also not acceptable.

Wald et al [314] have argued that a risk factor has to be extremely strongly associated with a disease within a population before it can be considered to be a potentially useful screening test. Even a odds ratio of 200 between the highest and lowest fifths will yield a detection rate of no more than about 56% for a 50% false positive rate. Another reason why strong risk factors may make poor screening tests, according to these researchers, is that there may be little variation in exposure within populations. They explain this by referring to the smoking example: It is known that smoking cigarettes is a risk factor for lung cancer. However, if everyone in a certain population smoked 20 cigarettes a day, asking about cigarette consumption would not distinguish those who are more likely to develop lung cancer from those who are not. In the present study the high prevalence of high fat intake could have had a similar effect.

As evident from Table 5.13 relatively few true negatives were found in all the comparisons between the screener and the measures of fat intake from the food record. This would contribute to the wide confidence intervals of the reported specificity (Table 5.14). Again the sample size would have played a role, as the required number of negatives necessary to yield the desirable power (see 4.1.1) was not obtained.

6.5.2.2 Receiver operating characteristic (ROC) curves

The high sensitivity but low specificity observed in this study was the rationale for investigating the effect of changing the cut-off value of the final score of the test method on the sensitivity and specificity relative to the various measures of high fat intake from the food record. By increasing the 'strictness' that is increasing the final score cut-off of the test method, the false positive rate should decrease, the sensitivity would also decrease with an associated increase in specificity.

ROC curves are useful to depict this pattern of sensitivities and specificities observed when the performance of the test method is evaluated at different cut-off values. They thus describe the whole set of (1-specificity, sensitivity), that is (false positive fraction, true positive fraction) combinations possible.[47]

In a ROC curve a test that perfectly discriminates between the two groups under discussion (that is 'high fat' versus 'prudent' diet consumers) would yield a curve that coincided with the left and top sides of the plot, that is having a high sensitivity (true positive rate) and a low false positive rate (1 minus specificity). Poor tests have lines close to the rising diagonal and consequently the area under the curve would be about 0.5. Hosmer and Lemeshow [315] have suggested that areas of 0.7 and higher can be taken as pointing to acceptable discrimination abilities of the test method. The shape of the curves obtained in this study showed that, regardless of which measure of high fat intake was used, the discrimination ability of the test method remained low, as the highest area under the curve was 0.65 (for PSFE). Equally, it was not possible to optimise the test method by manipulating the cut-off value of the final score.

This finding was in contrast to the Taylor et al study [205] where it was found that by substantially decreasing the cut-off value of the original MEDFICTS tool [203] a sensitivity and specificity of 0.73 and 0.75 respectively could be reached.

The results of the current study were more in line with those of Caan et al.[49] They reported that the sensitivity and specificity of a dietary fat screener varied depending on the cut-off point used, but it was not possible to achieve high sensitivity and high specificity simultaneously. Also Prochaska et al [316] found that their screening measure of fat intake was sensitive but not specific among adolescents.

In the development rationale of the test method (Chapter 3) the primary aim of the test method was stipulated as that it should be able to discriminate between children who consume high fat diets and those with prudent intakes. The ROC curve suggested that in the described context the dietary fat screener did not achieve this eventual goal.