Public Health Nutrition: 20(15), 2670–2679 doi:10.1017/S1368980017001471

Review Article

Developing and validating a nutrition knowledge questionnaire: key methods and considerations

Gina Louise Trakman[1,]*, Adrienne Forsyth[1], Russell Hoye[2] and Regina Belski[1]
[1]Department of Rehabilitation, Nutrition and Sport, School of Allied Health, College of Science, Health and Engineering, La Trobe University, Health Sciences 3 Building – Room 411, Cnr Plenty Road and Kingsbury Drive, Bundoora, VIC 3068, Australia: [2]Department of Management and Marketing, Centre for Sport and Social Impact, College of Arts, Social Sciences and Commerce Education, La Trobe University, Bundoora, Australia

Submitted 18 October 2016: Final revision received 20 April 2017: Accepted 26 May 2017: First published online 24 July 2017

Abstract

Objective: To outline key statistical considerations and detailed methodologies for the development and evaluation of a valid and reliable nutrition knowledge questionnaire.
Design: Literature on questionnaire development in a range of fields was reviewed and a set of evidence-based guidelines specific to the creation of a nutrition knowledge questionnaire have been developed. The recommendations describe key qualitative methods and statistical considerations, and include relevant examples from previous papers and existing nutrition knowledge questionnaires. Where details have been omitted for the sake of brevity, the reader has been directed to suitable references.
Results: We recommend an eight-step methodology for nutrition knowledge questionnaire development as follows: (i) definition of the construct and development of a test plan; (ii) generation of the item pool; (iii) choice of the scoring system and response format; (iv) assessment of content validity; (v) assessment of face validity; (vi) purification of the scale using item analysis, including item characteristics, difficulty and discrimination; (vii) evaluation of the scale including its factor structure and internal reliability, or Rasch analysis, including assessment of dimensionality and internal reliability; and (viii) gathering of data to re-examine the questionnaire's properties, assess temporal stability and confirm construct validity. Several of these methods have previously been overlooked.
Conclusions: The measurement of nutrition knowledge is an important consideration for individuals working in the nutrition field. Improved methods in the development of nutrition knowledge questionnaires, such as the use of factor analysis or Rasch analysis, will enable more confidence in reported measures of nutrition knowledge.

Keywords
Questionnaire
Nutrition knowledge
Factor analysis
Rasch analysis

Researchers employ nutrition knowledge questionnaires (NKQ) to benchmark levels of awareness of expert recommendations and to assess the effectiveness of nutrition education programmes using a pre-test/post-test method[1,2]. The development and validation of a questionnaire involves multiple complicated and time-consuming steps[2]; this procedure can be prohibitive and the appropriate procedures are often overlooked[3]. In fact, a 2002 review of evaluation measures used in nutrition education research (in pre-school children, school-aged children, adults and pregnant women) found that only 55 % of the studies in adults which used a questionnaire reported on the reliability of measures[3]. Likewise, a 2015 systematic review of sixty studies that used questionnaires to assess athletes' and coaches' nutrition attitudes and nutrition knowledge found that about 70 % of the included studies used tools of unknown validity and reliability, and 67 % used tools that had not undergone pilot testing. The authors of the review noted a number of issues related to statistical analysis, such as failure to report power calculations, confidence intervals and effect sizes[4]. Furthermore, there are issues with the content of the measures employed: a 2016 review of nutrition knowledge in athletes and coaches found that many tools based their questions on outdated recommendations and did not consider health literacy or cultural appropriation[5]. The use of poor-quality NKQ limits the conclusions that can be drawn from research on nutrition knowledge. This was

**Corresponding author:* Email g.trakman@latrobe.edu.au noted as early as 1985, in a meta-analysis by Axelson *et al.*[6] which reported on the correlation between nutrition knowledge and dietary intake. Similar conclusions were made in a 2014 review of the relationship between nutrition knowledge and diet quality, which reported a mean 'questionnaire quality score' of just 50 %[7].

Multiple journal articles and books have been published to provide guidelines for questionnaire development in the areas of behavioural psychology and management information systems[3,8–11], but there is a paucity of literature adapting this information to the development of an NKQ. To our knowledge, the only recommendations that exist are in the article by Parmenter and Wardle[2] in 2000, entitled 'Evaluation and design of nutrition knowledge measures'. These guidelines were employed to develop the widely used 'General Nutrition Knowledge Questionnaire' (GNKQ)[12] and have been followed by many other researchers developing nutrition knowledge measures[11,13,14]. Parmenter and Wardle[2] outline several techniques, based on Classical Test Theory (CTT), that are crucial for the psychometric validation of measurement tools. However, they do not recommend factor analysis, a technique that allows researchers to define potential 'factors' or nutrition sub-sections within their questionnaire[15]. They also make no mention of other frameworks for validation such as Item Response Theory (IRT) which includes Rasch analysis[16,17], an approach that allows researchers to develop shorter scales with multiple response formats. Since 2016 multiple nutrition knowledge questionnaires have been adapted from existing tools or developed and validated[18–26]; however, very few of these have undertaken factor analysis[22,24] or Rasch analysis[18,20].

The aim of the present review is to provide evidence-based recommendations for NKQ development and evaluation. The eight-step methodology (outlined in Box 1) integrates recommendations made by Parmenter and Wardle[2] and Pallant[27], and includes several crucial procedures from disciplines such as psychology and management information systems[8,28–30] that are frequently overlooked in nutrition. The review may provide guidance for researchers who are interested in developing new nutrition knowledge measures and/or evaluating the quality of existing tools.

Definitions and terminology

When reviewing the literature on questionnaire development, it is apparent that there are conflicting definitions for measurement properties related to reliability and validity. Throughout the present review, the definitions adopted are in line with those outlined in the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) taxonomy (Table 1). COSMIN was originally developed for the assessment of Health Related Patient Reported Outcomes (HR-PRO); readers are encouraged to refer to these guidelines as an accompanying resource to the present review[31].

Box 1. Outline of methods

Development of the tool:

1. Definition of the construct and development of a test plan.

2. Generation of the item pool.

3. Choice of the scoring system and response format.

Preliminary review of the items:

4. Assessment of content validity*.

5. Assessment of face validity*.

Further statistical analysis of measurement:

6. Purification of the scale using item analysis.

7. Evaluation of the scale including its (i) factor structure† and (ii) internal reliability†; OR Rasch analysis‡ including assessment of (i) dimensionality§ and (ii) internal reliability.

Final analysis:

8. Gathering of data to re-examine the questionnaire's properties, assess temporal stability‖ and confirm construct validity.

Notes

*These steps can be performed in reverse order.

†These steps are within a Classical Test Theory framework: item analysis includes item discrimination and item difficulty.

‡These steps are within an Item Response Theory framework and can also be performed after step 8.

§Dimensionality can be assessed in place of factor analysis.

‖Temporal stability can also be performed after step 6.

Key methodologies and statistical considerations

Step 1: Definition of the construct and development of a test plan

Researchers should begin questionnaire development by defining the construct that they intend to measure[2]. The definition should explain not only what the construct is, but also what it is not; knowledge should be distinguished from attitudes and behaviours. Developers may choose to adopt a generic definition, such as the one by Miller and Cassady[32]: 'Knowledge of concepts and processes related to nutrition and health including knowledge of diet and health, diet and disease, foods representing major sources of nutrients, and dietary guidelines and recommendations'. The exact topic of nutrition knowledge and the relevant nutrition sub-sections should be specified. Brown[29] refers to this as a 'test plan' and recommends that the relative importance (weighting) of each item is outlined. The test plan is likely to be quite diverse, depending on the intended purpose of the questionnaire.

Table 1 Definitions of psychometric measurement properties (adapted from the COSMIN taxonomy[31])

Term	Definition
Content validity	The ability of a questionnaire to adequately cover all relevant topics of the construct (concept) to be measured
Face validity	The degree to which the items of a questionnaire appear (on 'face value') as though they are an adequate reflection of the construct (concept) to be measured. Considered to be an aspect of content validity
Construct validity	The degree to which the scores on a questionnaire are consistent with hypotheses (e.g. with regard to differences between relevant groups) based on the assumption that the questionnaire validly measures the construct to be measured
Internal reliability	The degree of the interrelatedness among the items of a questionnaire; also referred to as internal consistency or homogeneity
Temporal stability	The ability of a questionnaire to detect change over time in the construct to be measured; also referred to external reliability
Responsiveness	The ability of a questionnaire instrument to detect change over time in the construct to be measured
Dimensionality*	The extent to which the items measure a hypothesized concept distinctly. In unidimensional scales, all items are said to reflect a single construct; in multidimensional scales, several topics (sub-sections) of the same construct are being measured
Criterion validity	The degree to which the scores of a questionnaire are an adequate reflection of a 'gold standard'

COSMIN, COnsensus-based Standards for the selection of health Measurement Instruments.
*This definition was not derived from COSMIN. In Classical Test Theory, dimensionality is determined by performing factor analysis.

For example, the GNKQ specified four nutrition sub-sections: dietary recommendations, sources of nutrients, choosing everyday foods and diet–disease relationship[12]. In contrast, the nutrition sub-sections of a questionnaire designed to assess type 1 diabetes nutrition knowledge were: healthful eating, carbohydrate counting, blood glucose response to foods and nutrition label reading[33].

Step 2: Generation of the item pool

Once the construct has been defined and the test plan made, a pool of items to represent each nutrition sub-section should be developed[2,30,34]. A certain degree of redundancy in questions is recommended, so that questions that do not behave as expected can be removed at later validation stages[30]. DeVellis[30] suggests that the number of items included in the first draft of the questionnaire should be up to three to four times the amount that will appear in the finalized version. Parmenter and Wardle[2] recommend writing twice as many questions as you want to have in your final tool.

The creation of new questionnaire items should be guided by expert opinion and the current literature including peer-reviewed journal articles and education materials available to the public. Items can also be taken from previous questionnaires, either in their original form or modified to suit the purpose of the research[2]. If items from previous questionnaires are used, permission should be sought and credit given to the original authors[2,35].

The language used should be kept as simple and concise as possible. Double negatives and two-edged questions ('a diet high in fruits and vegetables AND low in salt can help prevent high blood pressure'[36]) should be avoided[1,34,35,37] because they tend to be ambiguous and confuse respondents. Questions should be written as full sentences, and slang and abbreviations should not be used. Jargon and technical terms can be used with caution, provided the group being assessed is expected to be familiar with these terms[35]. In some instances, it is recommended that interviews with the target audience be conducted so that their vernacular can be accurately captured[37]. The names used for foods must be commonly understood and relevant for the target audience. Where previous items are used, it may be necessary to make language adjustments. For example, when modifying the GNKQ (developed for a UK audience) to be used with an Australian sample, Hendrie *et al.*[38] used the term '35 % orange juice' instead of 'orange squash'. Similarly, when validating the GNKQ in a Turkish sample, Alsaffar[14] replaced 'baked beans on toast' with 'piyaz', which is a white bean salad commonly eaten in Turkey.

Step 3: Choice of the response format and scoring system

This step should be conducted simultaneously to writing the questionnaire items. There is no ideal response format and scoring system; the relative pros and cons of various options are outlined below and should be considered in relation to the specific purpose of the questionnaire that is being developed.

Response format

The first decision to make is whether open-ended (participants provide responses) or close-ended (pre-selected responses) will be used. The former are more difficult for respondents to answer and for researchers to code, and therefore are regarded as less reliable[35,39]. The main benefit of open-ended questions lies in their ability to capture unexpected answers, for example when quotes or testimonies are required; they rarely provide an advantage where the aim is to measure nutrition knowledge[35]. A review of current nutrition measures revealed that with the exception of the diet–disease relationship section of the GNKQ[12], open-ended questionnaires are not used.

For close-ended questions, possible response formats include true/false, yes/no, Likert scale (e.g. 'strongly agree', 'agree', 'disagree', 'strongly disagree') and multiple-choice (usually with four to five options)[35]. Agree/disagree-type scales tend to reduce the feeling by participants that they are being tested and judged[1]. Multiple-choice options are useful because the analysis of distractor options can provide valuable information regarding nutrition misinformation[17]. It is not uncommon for participants to be able to select several options (e.g. 'assuming equal weights please choose 10 foods that you think are high in fiber'[40]). In general, 'select all options that are correct' questions can be difficult to 'code' and score, and should be avoided where possible. Several authors who have developed an NKQ have employed a 'not sure' or equivalent category[12,14,38,40–44]; however, many have chosen not to provide this option[45–48]. The benefit of a 'not sure' option is that it may prevent respondents from correctly guessing the correct option, the chances of which are 50 % for dichotomous items[1]. On the other hand, this category may provoke laziness, or lead those who have a good idea of responses to avoid answering if confidence is low[2]. A range of question styles and responses is likely to be suitable. Sudman and Bradburn[1] even recommend that some pictorial questions are included to avoid monotony and reduce respondent fatigue.

Scoring system

The most common scoring system is to simply award a point for each time the correct option is selected[12,14,40]. Negative scoring can also be used (e.g. Zawila *et al.*[43] awarded 1 point for correct, 0 points for 'not sure' and deducted 1 point for incorrect options). Likert scales can be challenging to score; Hoogenboom *et al.*[49] grouped 'strongly agree' and 'agree', and 'strongly disagree' and 'strongly disagree', and awarded 1 point where true statements were endorsed or false statements were renounced. Sedek and Yih[50] scored each question with 1 to 4 points, with 4 points being given if a true statement was strongly agreed with, 3 points for agree, and so on. Researchers should also consider whether a total summed score is appropriate or not. It has been suggested that total scores are appropriate only where the construct has been proved to be unidimensional[2]. Subscores are likely to be important when gaps in knowledge need to be assessed for education purposes. Appropriate methodology for accurately assessing the dimensionality of measures is described in Box 3 and step 7.

The order in which questions are to be asked needs to be carefully considered. Ideally, the answer to one question should not be able to be ascertained from a preceding question[1,2]. A common recommendation is to start with easy, necessary non-threatening questions and to avoid asking demographic questions at the beginning if possible, because these can be seen as probing and therefore off-putting[39].

Step 4: Assessment of content validity

Once the items have been developed and the appropriate response formats determined, the questions should be reviewed by a panel of experts[1,8,34]. In the case of an NKQ, these should be dietitians or nutritionists, preferably working in a range of areas such as academia, private practice and industry. It may also be appropriate to include individuals with expertise in survey design. The aim of this step is to ensure that the tool has adequate content validity. That is, that questions being posed are relevant and cover all topics of the 'construct' as defined in step 1[51].

Several researchers who have developed questionnaires state an expert panel review/review for content validity was performed[11,12,42,52,53], but they do not describe the way in which data were collected or analysed. It appears that qualitative data were collected in an *ad hoc* manner.

A broader search of the literature reveals that content validity can be quantified using several methods, such as the content validity index (CVI)[51]. In order to calculate the CVI, a group of three to ten experts is required. Each expert rates individual items for relevance on a 4-point Likert scale (1 = 'very irrelevant', 2 = 'irrelevant', 3 = 'relevant', 4 = 'very relevant'). The CVI for each question is calculated by dividing the number of raters who scored the item as 3 or 4 divided by the total number of raters; a score above 0·8 is considered adequate. Ratings for accuracy, clarity and appropriateness can also be obtained and analysed qualitatively[51,54].

MacKenzie *et al.*[34] propose an alternative method for assessing content validity. This involves constructing a matrix with definitions of each nutrition sub-section the questionnaire is aiming to test listed along the top and the questionnaire items listed down the side. Each 'rater' then indicates on a 4- to 5-point Likert scale how well the item captures each sub-section (Table 2). Repeated one-way ANOVA can be used to assess if an item's rating on each topic differs significantly; items should score higher on the topic they intend to assess (as per the test plan). This approach is appropriate only when fewer than eight to ten nutrition sub-sections are being assessed, allowing for the inclusion of some distractors (i.e. some nutrition sub-sections that are not being tested should be included in the matrix).

Step 5: Pre-testing and assessment of face validity

In addition to having the items reviewed by a panel of experts, a small sample of the target audience (ten to twenty participants) should also complete the questionnaire before recruiting the final sample[34]. This step: (i) confirms that the instructions given are easy to follow and there are no technical issues with completing the tool (especially important if it is in an online format); (ii) gives an indication of how long the questionnaire will take to complete[30]; and (iii) allows face validity[31] to be assessed.

As with content validity, the specifics of how pre-testing and face validity assessment have been conducted by researchers are unclear. In general, it is simply stated that

Table 2 A hypothetical example of an item rating task to assess content adequacy of a questionnaire (inspired by MacKenzie *et al.*[34])

Rater = 001 Sports Nutrition Knowledge Questionnaire (true/false statements)	Hydration	Recovery	Supplementation
For optimal recovery, athletes should drink 0·5 g of fluid for every 1·0 g of weight that is lost during training or competition	4	3	1
Creatine supplements would be most beneficial to a player wanting to increase peak power output during repeated bouts of exercise	1	1	4
Fluid consumed for hydration purposes (during exercise) should contain at least 4–8 mmol/l (~90–185 g/l) of sodium (salt)	5	3	2

Numbers represent the perceived extent to which each item captures each aspect of the construct domain using a 5-point Likert-type scale ranging from 1 ('not at all') to 5 ('completely').

feedback on topics such as clarity and understanding of items was obtained[10,55].

A reliable technique used for face validity assessment is the think-out-loud model, whereby participants verbalize their thought process as they complete each item[17]. This can be done retrospectively; that is, the participant can complete the questionnaire in advance and then meet with the researcher to discuss his/her experience of completing the questionnaire[56]. A less formal approach may be conducting a focus group where questions such as 'what do you think this section is testing?', 'are you unfamiliar with any of the terms used in this question?', 'do you find this question confusing or intentionally misleading?' can be asked, with responses being analysed to make necessary changes to terminology and wording of items.

It is not usually essential to redo face validity testing; however, this may be necessary if the sample being recruited is quite different from the cohort on whom the original test was validated. Many authors who have used pre-existing questionnaires have repeated these steps[57,58].

Step 6: Purification and refinement using item analyses

Once the items have been updated to ensure content and face validity, it is necessary to recruit another sample similar to the target sample to perform item analyses[34]. Item analyses refers to a range of CTT techniques, including assessment of item characteristics, item difficulty and item discrimination[2,8,9,30,59]. The general features of CTT are covered in Box 2.

Box 2 Classical Test Theory: premise and sample size recommendations

- The underlying premise of Classical Test Theory (CTT) is that that a person's true score on a measure is a function of their observed score and measurement error[17].
- Mathematically, CTT is based on correlations between items. Validation using CTT only applies to the group of people who were used to assess the tool; scales validated using these techniques need to be reassessed every time they are used[60].
- DeVellis[30] suggests a sample size of 100 is 'poor', 200 is 'fair' and 300 is 'good'; Parmenter and Wardle[2] advise that the number of respondents should be at least one greater than the number of questions. McCoach *et al.*[9] recommend six to ten times as many respondents as questions. These figures are often not achieved by researchers developing nutrition knowledge measures[11,14,42]. MacKenzie *et al.*[34] note that if the correlation between items is a high, smaller sample sizes are likely to be appropriate.

Item characteristics

Gable and Wolf[61] suggest that response frequencies, means and standard deviations should be 'screened' and that researchers should consider removing items whose responses are very positively or negatively skewed (see below for recommended cut-offs)[27,45]. For multiple-choice questions, the frequency with which incorrect options are chosen should be assessed; Petrillo *et al.*[62] state options chosen by less than 5 % of participants are 'non-functional distractors' and should be modified. Parmenter and Wardle[2] recommend all distractors should be endorsed by an equal number of respondents. Assessment of distractor options does not appear to have been undertaken (or at least not reported) in the nutrition knowledge field[2].

Item difficulty

Item difficulty (sometimes called item facility[2] or item severity[63]) should be assessed by reviewing how frequently respondents answered individual questions correctly. If less than 20 % or more than 80 % of respondents answered an item correctly, its removal should be considered[8,12]. Many researchers who have evaluated nutrition knowledge measures have removed items on this basis[11,12,41,64,65]. However, individual questions may have utility beyond their contribution to the total knowledge score. For example, observing that responses to a question are consistently poor provides valuable information about gaps in knowledge. Therefore, researchers must employ pragmatic decision-making processes before deleting or modifying items[30].

Item discrimination

If a person does well overall but poorly on a particular item (and vice versa), the item is said to be a poor judge (or discriminator) of knowledge[2]. Item discrimination can be assessed based on the correlation between an item and the total score (minus the item of interest). Minimum correlation coefficients of 0·2–0·3 are recommended[2,59]. Pearson's correlation coefficient should be used if questions are multi-choice, and point-biserial correlation, which is a special case of Pearson's *r*, should be used if items are dichotomous[66]; in previous literature, either the distinction between these statistics has not been made[2] or it is unclear what correlation coefficient has been used[64,67,68]. A second method for item discrimination, described by Cappelleri *et al.*[63], is to divide respondents into high scoring and low scoring groups (using cut-offs such as 25th and 75th percentile) and to then evaluate the percentage of individuals in each group who endorsed correct/incorrect statements.

Inter-item correlations

The inter-item correlations among scale items can also be evaluated[61]. It is said that items with very high correlations (r = 0·7) may be measuring the same thing, whereas items with low correlations (r = 0·3) may reflect items that are too diverse to be assessing a single construct[27,28]. Assessment of inter-item correlations does not appear to have been performed in previous studies that have used psychometrics to validate nutrition knowledge measures. This may be because in an achievement test (as opposed to a personality type test) more items may be required to assess certain constructs such as knowledge of hydration, yet knowledge of these items would be expected to be highly correlated. We recommend that it is still worthwhile assessing inter-item correlations; however, item pairs with high correlations should be assessed qualitatively to decide if they are redundant before removing one of them.

Step 7a: Evaluation of the scale's factor structure (using exploratory factor analysis) and internal reliability

Exploratory factor analysis (EFA) is a CTT technique that allows for a mathematical 'exploration' of the number of variables or 'factors' within a scale[28,59]. EFA provides information regarding the underlying dimensionality of the measure[15,27]. The most commonly used technique to assess factor structure is principal components analysis (PCA)[27,28]. In CTT, data must be assessed to ensure several conditions are met before proceeding with PCA. These include an adequate sample size, inter-item correlations of at least 0·3, a significant ($P \leq$ 0·05) Bartlett's test of sphericity, and a Kaiser–Meyer–Oklin (KMO) measure of sampling adequacy (MSA) of at least 0·6[70]. A common assumption for using PCA is that variables must be continuous[71]. However, if variables are dichotomous (e.g. right/wrong), which is often the case with NKQ, factor analysis can still be conducted, provided tetrachoric rather than Pearson's *r* correlations are used[71].

Box 3 Factor structure and dimensionality

For a measurement tool to be described as unidimensional, all of its items must be measuring the same underlying construct[28]. In the nutrition knowledge field, the 'dimensions' in a scale are usually defined by the questionnaire developer *a priori*. For example, expert opinion is used to decide that items one to four assess knowledge of government-endorsed 'dietary recommendations' and hang together in their own sub-section[12]. However, there is minimal evidence of authors undertaking a formal assessment of factor structure using Classical Test Theory (CTT)[24,68,69]. An outline of how to undertake factor analysis using CTT is described in step 7a. An alternative is to use Rasch analysis to confirm unidimensionality[18,20]. Additional detail on Rasch analysis is outlined in Box 4 and step 7b.

Deciding how many factors your tool has is both an art and a science, and often a single solution does not exist. In CTT, the final decisions should be based on results of a variety of tests including Kaiser's criterion, Cartel's scree plot and percentage variance[27,70]. If a measure is found to have more than one factor, the researcher can rotate the factor solution to assist in deciphering which items belong to which factor. Factors that are rotated are often less ambiguous and therefore easier to interpret[72].

Lin and Ya-Wen[69] performed factor analysis when developing an NKQ for use in elderly Taiwanese and found that the questionnaire had three subscales: nutrition and disease; requirements of food groups; and nutrients in food. Bradette-Laplante *et al.*[24] conducted EFA on an NKQ developed for a Canadian population and found that the tool had only one factor. Guadagnin *et al.*[22] developed a questionnaire to assess the effectiveness of a nutrition education programme in the workplace and found it had four factors.

For more detailed information on performing factor analysis using CTT, the reader is directed to other publications[28,71–74].

Internal reliability in Classical Test Theory

Internal reliability assesses the degree to which items within a questionnaire measure different topics of the same construct. A high measure of internal reliability is said to be reflective of a small degree of random error[8]. Internal reliability is considered one of the most important determinants of reliability within the CTT framework[30]. It is assessed using Kruder–Richardson (KR-20) for dichotomous scales and Cronbach's alpha (Cα) for items with more than two responses(60). Both statistics range from 0 to 1, with 1 indicating perfect correlation; a cut-off point of 0·7 is frequently cited as adequate(2,59). It is important to consider the limitations of these statistics. These include:

- Not appropriate for multidimensional questionnaires. Cα and KR-20 presume that all items are an equal measure of the underlying construct. Therefore, they should be used only if a scale is found to be unidimensional, or should be used only on individual sub-sections of an NKQ(60). In NKQ, it is common for Cα to be reported for individual nutrition sub-sections (that assess a particular topic of knowledge) rather than the scale as a whole. A good option may also be to run Cα or KR-20 according to factor structure, after factor analysis has been performed.
- Not appropriate for longer questionnaires. Long questionnaires will automatically achieve a higher Cα(8). Therefore, to avoid a falsely 'respectable' Cα or KR-20, Streiner(60) recommends that it should not be used on scales that have more than twenty items.
- Values are difficult to interpret. While some authors state very high Cα ($r = 0{\cdot}9$) is favourable(59), others argue that these values may point to redundancy in items and that inter-item correlations are, in fact, a better measure of reliability(28).

Step 7b: Rasch analysis to evaluate items including assessment of (i) dimensionality (ii) and internal reliability

The general characteristics of IRT and Rash analysis are covered in Box 4.

Rasch analysis aims to produce scales that are unidimensional; multidimensionality can result in misfit to the Rasch model(75). During Rasch analysis, dimensionality can be assessed by conducting a PCA on 'residuals'. Residuals are calculated based on the differences between observed and expected data(17). PCA on residuals results in identification of sub-sets of items, which can then be assessed for similarity to/differences from each other using a *t* test(75). In Rasch analysis internal reliability can be evaluated using the person separation index (PSI). The PSI is evaluated using the same criteria as Cα, with a value of 0·7 said to be indicative of adequate internal reliability(75).

Rasch analysis has been used to validate a scale on clinical nutrition literacy(76). Likewise, Motteli *et al.*(20) used Rasch analysis to develop scales on practical knowledge of balanced meals(20) and the energy content of meals(18). The authors noted that in their studies, the ability of Rasch scaling to separately estimate item difficulty and person ability, and to develop brief tools, offered a major advantage over CTT.

Box 4 Item Response Theory: premise and sample sizes

- The underlying premise of Item Response Theory (IRT) is that the probability of an individual answering a particular item correctly is dependent on the underlying level of the construct being measured (e.g. his/her level of nutrition knowledge and the difficulty of the item)(17). This is represented graphically, using item characteristic curves (ICC) for dichotomous items or category response curves (CRC) for multiple-choice and Likert-scale items(63).
- Rasch analysis is one of the most commonly used IRT models(17). Rasch analysis involves performing multiple different types of 'diagnostics' before making a final decision on which items and response formats to modify and whether to delete certain persons from the analysis or not(63).
- Scales that are shown to fit the Rasch model are said to be inherently valid(17); this is because IRT does not rely on measures of central tendency which are influenced by the sample characteristics.
- Sample sizes for IRT vary widely, but are smaller for Rasch models compared with other IRT techniques. For models with dichotomous items (which nutrition knowledge measures are likely to be), 200 respondents is thought to be adequate(18,75).

Detailed instructions regarding how to complete PCA on residuals, and Rasch analysis in general, are beyond the scope of the current review. Largely because the program used (RUMM, WINSTEPS, POLM, MULTILOG, PAESCALE, BILOG or NLMIXED) has an effect on the statistics (also referred to as 'indicators') produced(75), it is likely that specialized training will be needed to complete this type of analysis. For additional information on Rasch analysis, the reader is directed to Tennant and Conaghan(75), Pallant and Tennant(77), Pallant(27) and Presser *et al.*(17). A table outlining the definition and interpretation of key statistics produced during Rasch analysis, with examples relevant to the RUMM program, is available in the online supplementary material, Supplemental Table 1.

It is not necessary to conduct factor analysis AND Rasch analysis; scales that are found to have multiple factors are likely to be multidimensional and misfit the Rasch model(78). However, item analysis (step 6, within the CTT framework) and Rasch analysis (step 7b, within the IRT framework) can be conducted on the same scale. Readers may choose to review examples of scales that have been validated using both CTT and IRT, such as Fan(16). If this approach is taken, indicators from each framework will need to be compared and contrasted before making decisions about which items to delete, modify and keep.

Step 8: Gather data to re-examine and assess validity using known-group comparisons/cross-validate the scale

Once steps 1 to 6 have been performed, you may find you have a questionnaire that is considerably different from the one with which you started. Therefore, it is preferable to re-administer the tool to a new sample, so that item analysis and reliability can be re-evaluated[34]. If it has not yet been performed, factor analysis or Rasch analysis can also be attempted at this time. There are examples in other disciplines of scales that have been used for many years before being analysed to assess if they meet the Rasch model[77,79]. This 'second pilot' can be used to assess construct validity and temporal stability, although these could also technically be conducted at the same time (on the same sample) as item analysis.

Construct validity

Construct validity can be assessed by performing known-group comparisons[31]; that is, by statistically comparing mean scores of groups that you expect to do well with mean scores from participants whose knowledge should not be as high. This serves to confirm the test is measuring what you think it is[2]. Previous researchers developing nutrition knowledge measures have done this by comparing: home economics and other students, university dietetic and non-dietetic students[11], university students studying nutrition with individuals without nutrition education[38], dietetic students *v.* nursing interns[64], individuals with *v.* without nutrition qualifications[80], and nutrition experts *v.* computer experts[2]. Previous researchers have found that females' knowledge is greater than males' knowledge and that age (middle age > young > elderly) can have an effect on nutrition knowledge[12]. However, some caution needs to be taken if using these relationships to assess construct validity, because there is some conflict regarding whether they always occur[5].

Temporal stability

Temporal stability (also referred to as external reliability in the nutrition knowledge field)[2] can be assessed using the test–retest method; that is, by administrating the test on two separate occasions and assessing the correlations between individuals' scores on the two attempts, using Pearson's *r*. Correlation should be about 0·7[2]. The time between test attempts should be long enough that exact answers provided are forgotten, but short enough that no new information is learnt; two weeks are commonly used[2,42,81,82]. A limitation of this method is that motivated individuals may look up the answers to questions they answered incorrectly and thereby increase their knowledge between test occasions.

Additional types of validation that may be relevant are outlined in Box 5.

Conclusions and implications for practice

In conclusion, the measurement of nutrition knowledge is an important consideration for individuals working in the nutrition field. We have outlined key methodologies and considerations for researchers who are interested in developing or evaluating a nutrition knowledge measure. Many published NKQ fail to adequately describe how content and face validity were assessed, have not undergone assessment of distractor utility, and have not had their dimensionality assessed. Improved methods used in the development of NKQ will enable more confidence in reported measures of nutrition knowledge. The authors recommend that all new measures to assess nutrition knowledge consider the methodologies described in the present review. Likewise, it is recommended that existing scales undergo factor analysis or Rasch analysis to confirm their dimensionality, reliability and validity.

Box 5 Occasionally recommended supplementary validation

Criterion validity

If a gold standard tool for measuring nutrition knowledge exists, a new tool is unlikely to be developed. Therefore, in the field, this type of validity is likely to be required only if a shortened form of an existing validated questionnaire is being developed[80].

Responsiveness

Researchers who are intending on using their questionnaire to assess changes in knowledge over time, or before and after an education programme, should also validate their tool for responsiveness. Responsiveness is most commonly measured by developing a hypothesis that outlines expected changes and then administering your tool alongside a gold standard at two time points to test the hypothesis[31]. Responsiveness testing is rare in previous nutrition knowledge measures; however, it was conducted in validation of the recently revised GNKQ[21]. For more information on responsiveness in health measures, readers are directed to Hays *et al.*[83].

Acknowledgements

Acknowledgements: The authors would like to acknowledge Julie Pallant, whose tutelage at the 2016 ASCPRI (Australian Consortium for Social & Political Research Inc.) course, 'Scale development, Rasch Analysis and Item Response Theory', was invaluable in building our understanding of questionnaire development. *Financial support:* This work was supported by an Australian Government Research Training Program Scholarship. *Conflict of interest:* None. *Authorship:* G.L.T. developed the guidelines. All authors contributed to writing and revising the paper for intellectual content. *Ethics of human subject participation:* Not applicable.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S1368980017001471

References

1. Sudman S & Bradburn NM (1982) *Asking Questions: A Practical Guide to Questionnaire Design*. New York: John Wiley & Sons, Inc.
2. Parmenter K & Wardle J (2000) Evaluation and design of nutrition knowledge measures. *J Nutr Educ* **32**, 269–277.
3. Frary RB (2003) A brief guide to questionnaire development. Virginia Polytechnic Institute State University. http://medrescon.tripod.com/questionnaire.pdf (accessed June 2015).
4. Kouvelioti R & Vagenas G (2015) Methodological and statistical quality in research evaluating nutritional attitudes in sports. *Int J Sport Nutr Exerc Metab* **25**, 624–635.
5. Trakman G, Forsyth A, Devlin B *et al.* (2016) A systematic review of athletes' and coaches' nutrition knowledge and reflections on the quality of current nutrition knowledge measures. *Nutrients* **8**, 570–593.
6. Axelson ML, Federline TL & Brinberg D (1985) A meta-analysis of food-and nutrition-related research. *J Nutr Educ* **17**, 51–54.
7. Spronk I, Kullen C, Burdon C *et al.* (2014) Relationship between nutrition knowledge and dietary intake. *Br J Nutr* **111**, 1713–1726.
8. Kline P (2013) *Handbook of Psychological Testing*. Abingdon: Routledge.
9. McCoach DB, Gable RK & Madura JP (2013) *Instrument Development in the Affective Domain*. New York: Springer.
10. Venter I (2008) Construction of a valid and reliable test to determine knowledge on dietary fat of higher-educated young adults. *S Afr J Clin Nutr* **21**, 133–139.
11. Whati L, Senekal M, Steyn N *et al.* (2005) Development of a reliable and valid nutritional knowledge questionnaire for urban South African adolescents. *Nutrition* **21**, 76–85.
12. Parmenter K & Wardle J (1999) Development of a general nutrition knowledge questionnaire for adults. *Eur J Clin Nutr* **53**, 298–308.
13. De Souza RS, Kratzenstein S, Hain G *et al.* (2015) General Nutrition Knowledge Questionnaire – modified and validated for use in German adolescent athletes. *Dtsch Z Sportmed* **66**, 248–252.
14. Alsaffar AA (2012) Validation of a general nutrition knowledge questionnaire in a Turkish student sample. *Public Health Nutr* **15**, 2074–2085.
15. Kim J-O & Mueller CW (1978) *Introduction to Factor Analysis: What It Is and How to Do It*. New York: SAGE Publications, Inc.
16. Fan X (1998) Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educ Psychol Meas* **58**, 357–381.
17. Presser S, Couper MP, Lessler JT *et al.* (2004) Methods for testing and evaluating survey questions. *Public Opin Q* **68**, 109–130.
18. Mötteli S, Barbey J, Keller C *et al.* (2017) Development and validation of a brief instrument to measure knowledge about the energy content of meals. *J Nutr Educ Behav* **49**, 257–263.e1.
19. Furber MJW, Roberts JD & Roberts MG (2017) A valid and reliable nutrition knowledge questionnaire for track and field athletes. *BMC Nutr* **3**, 36–43.
20. Motteli S, Barbey J, Keller C *et al.* (2016) Measuring practical knowledge about balanced meals: development and validation of the brief PKB-7 scale. *Eur J Clin Nutr* **70**, 505–510.
21. Kliemann N, Wardle J, Johnson F *et al.* (2016) Reliability and validity of a revised version of the General Nutrition Knowledge Questionnaire. *Eur J Clin Nutr* **70**, 1174–1180.
22. Guadagnin SC, Nakano EY, Dutra ES *et al.* (2016) Workplace nutrition knowledge questionnaire: psychometric validation and application. *Br J Nutr* **116**, 1546–1552.
23. Bukenya R, Ahmed A, Chapman-Novakofski KM *et al.* (2016) Validation of general nutrition knowledge questionnaire for adults in Uganda. *FASEB J* **30**, 896.13.
24. Bradette-Laplante M, Carbonneau É, Provencher V *et al.* (2017) Development and validation of a nutrition knowledge questionnaire for a Canadian population. *Public Health Nutr* **20**, 1184–1192.
25. Oz F, Aydin R, Onsuz MF *et al.* (2016) Development of a reliable and valid adolescence nutritional knowledge questionnaire. *Prog Nutr* **18**, 125–134.
26. Bottcher M, Marincic P, Nahay K *et al.* (2016) Nutrition knowledge and Mediterranean diet adherence: validation of a field based survey instrument. *J Acad Nutr Diet* **116**, 166–176.
27. Pallant J (2016) Scale development, Rasch Analysis and Item Response Theory [Course Notes] (In the Press).
28. Briggs SR & Cheek JM (1986) The role of factor analysis in the development and evaluation of personality scales. *J Pers* **54**, 106–148.
29. Brown FG (1983) *Principles of Educational and Psychological Testing*. Belmont, CA: Wadsworth Publishing Co.
30. DeVellis RF (2016) *Scale Development: Theory and Applications*. New York: SAGE Publications, Inc.
31. Mokkink LB, Terwee CB, Patrick DL *et al.* (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* **63**, 737–745.
32. Miller LMS & Cassady DL (2015) The effects of nutrition knowledge on food label use. A review of the literature. *Appetite* **92**, 207–216.
33. Rovner AJ, Nansel TR, Mehta SN *et al.* (2012) Development and validation of the type 1 diabetes nutrition knowledge survey. *Diabetes Care* **35**, 1643–1647.
34. MacKenzie SB, Podsakoff PM & Podsakoff NP (2011) Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. *MIS Q* **35**, 293–334.
35. Fink A (2002) *How to Ask Survey Questions*. New York: SAGE Publications, Inc.
36. McLeod ER, Campbell KJ & Hesketh KD (2011) Nutrition knowledge: a mediator between socioeconomic position and diet quality in Australian first-time mothers. *J Am Diet Assoc* **111**, 696–704.
37. Dawis RV (1987) Scale construction. *J Couns Psychol* **34**, 481–489.
38. Hendrie GA, Cox DN & Coveney J (2008) Validation of the general nutrition knowledge questionnaire in an Australian community sample. *Nutr Diet* **65**, 72–77.
39. Rattray J & Jones MC (2007) Essential elements of questionnaire design and development. *J Clin Nurs* **16**, 234–243.
40. Shepherd R & Towler G (1992) Nutrition knowledge, attitudes and fat intake: application of the theory of reasoned action. *J Hum Nutr Diet* **5**, 387–397.
41. Worme JD, Doubt TJ, Singh A *et al.* (1990) Dietary patterns, gastrointestinal complaints, and nutrition knowledge of recreational triathletes. *Am J Clin Nutr* **51**, 690–697.
42. Zinn C, Schofield G & Wall C (2005) Development of a psychometrically valid and reliable sports nutrition knowledge questionnaire. *J Sci Med Sport* **8**, 346–351.
43. Zawila LG, Steib CM & Hoogenboom B (2003) The female collegiate cross-country runner: nutritional knowledge and attitudes. *J Athl Train* **38**, 67–74.

44. Fanelli MT & Abemethy MM (1986) A nutritional questionnaire for older adults. *Gerontologist* **26**, 192–197.
45. Collison SB, Kuczmarski MF & Vickery CE (1996) Impact of nutrition education on female athletes. *Am J Health Behav* **20**, 14–23.
46. Corley G, Demarest-Litchford M & Bazzarre TL (1990) Nutrition knowledge and dietary practices of college coaches. *J Am Diet Assoc* **90**, 705–709.
47. Nichols PE, Jonnalagadda SS, Rosenbloom CA *et al.* (2005) Knowledge, attitudes, and behaviors regarding hydration and fluid replacement of collegiate athletes. *Int J Sport Nutr Exerc Metab* **15**, 515–527.
48. Folasire OF, Akomolafe AA & Sanusi RA (2015) Does nutrition knowledge and practice of athletes translate to enhanced athletic performance? Cross-sectional study amongst Nigerian undergraduate athletes. *Glob J Health Sci* **7**, 215–225.
49. Hoogenboom BJ, Morris J, Morris C *et al.* (2009) Nutritional knowledge and eating behaviors of female, collegiate swimmers. *N Am J Sports Phys Ther* **4**, 139–148.
50. Sedek R & Yih TY (2014) Dietary habits and nutrition knowledge among athletes and non-athletes in National University of Malaysia (UKM). *Pak J Nutr* **13**, 752–759.
51. Polit DF, Beck CT & Owen SV (2007) Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health* **30**, 459–467.
52. Feren A, Torheim LE & Lillegaard IT (2011) Development of a nutrition knowledge questionnaire for obese adults. *Food Nutr Res* **2011**, 55.
53. Jones AM, Lamp C, Neelon M *et al.* (2015) Reliability and validity of nutrition knowledge questionnaire for adults. *J Nutr Educ Behav* **47**, 69–74.
54. Wynd CA, Schmidt B & Schaefer MA (2003) Two quantitative approaches for estimating content validity. *West J Nurs Res* **25**, 508–518.
55. Raymond-Barker P, Petroczi A & Quested E (2007) Assessment of nutritional knowledge in female athletes susceptible to the female athlete triad syndrome. *J Occup Med Toxicol* **2**, 10–21.
56. Drennan J (2003) Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *J Adv Nurs* **42**, 57–63.
57. Alaunyte I, Perry JL & Aubrey T (2015) Nutritional knowledge and eating habits of professional rugby league players: does knowledge translate into practice? *J Int Soc Sports Nutr* **12**, 1.
58. Spendlove JK, Heaney SE, Gifford JA *et al.* (2012) Evaluation of general nutrition knowledge in elite Australian athletes. *Br J Nutr* **107**, 1871–1880.
59. Nunnally J (1978) *Psychometric Methods*. New York: McGraw-Hill.
60. Streiner DL (2003) Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess* **80**, 99–103.
61. Gable RK & Wolf MB (2012) *Instrument Development in the Affective Domain: Measuring Attitudes and Values in Corporate and School Settings*. Berlin: Springer Science & Business Media.
62. Petrillo J, Cano SJ, McLeod LD *et al.* (2015) Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health* **18**, 25–34.
63. Cappelleri JC, Lundy JJ & Hays RD (2014) Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clin Ther* **36**, 648–662.
64. Steyn NP, Labadarios D, Nel JH *et al.* (2005) Development and validation of a questionnaire to test knowledge and practices of dietitians regarding dietary supplements. *Nutrition* **21**, 51–58.
65. Shifflett B, Timm C & Kahanov L (2002) Understanding of athletes' nutritional needs among athletes, coaches, and athletic trainers. *Res Q Exerc Sport* **73**, 357–362.
66. Meyer JP (2014) *Applied Measurement with jMetrik*. Abingdon: Routledge.
67. Glanz K, Kristal AR, Sorensen G *et al.* (1993) Development and validation of measures of psychosocial factors influencing fat- and fiber-related dietary behavior. *Prev Med* **22**, 373–387.
68. Dwyer JT, Stolurow KA & Orr R (1981) A nutrition knowledge test for high school students. *J Nutr Educ* **13**, 93–94.
69. Lin W & Ya-Wen L (2005) Nutrition knowledge, attitudes, and dietary restriction behavior of the Taiwanese elderly. *Asia Pac J Clin Nutr* **14**, 221–229.
70. Pallant J (2013) *SPSS Survival Manual*. Sydney: Allen & Unwin.
71. Woods CM (2002) Factor analysis of scales composed of binary items: illustration with the Maudsley Obsessional Compulsive Inventory. *J Psychopathol Behav Assess* **24**, 215–223.
72. Yong AG & Pearce S (2013) A beginner's guide to factor analysis: focusing on exploratory factor analysis. *Tutor Quant Method Psychol* **9**, 79–94.
73. Bartholomew DJ (1980) Factor analysis for categorical data. *J R Stat Soc Ser B* **42**, 293–321.
74. Mislevy RJ (1986) Recent developments in the factor analysis of categorical variables. *J Educ Behav Stat* **11**, 3–31.
75. Tennant A & Conaghan PG (2007) The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res (Hoboken)* **57**, 1358–1362.
76. Guttersrud O, Dalane JO & Pettersen S (2014) Improving measurement in nutrition literacy research using Rasch modelling: examining construct validity of stage-specific 'critical nutrition literacy' scales. *Public Health Nutr* **17**, 877–883.
77. Pallant JF & Tennant A (2007) An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* **46**, 1–18.
78. Al-shair K, Kolsum U, Berry P *et al.* (2009) Development, dimensions, reliability and validity of the novel Manchester COPD fatigue scale. *Thorax* **64**, 950–955.
79. Hagquist C, Bruce M & Gustavsson JP (2009) Using the Rasch model in nursing research: an introduction and illustrative example. *Int J Nurs Stud* **46**, 380–393.
80. Dickson-Spillmann M, Siegrist M & Keller C (2011) Development and validation of a short, consumer-oriented nutrition knowledge questionnaire. *Appetite* **56**, 617–620.
81. Abood DA, Black DR & Birnbaum RD (2004) Nutrition education intervention for college female athletes. *J Nutr Educ Behav* **36**, 135–137.
82. Shoaf LR, McClellan PD & Birskovich KA (1986) Nutrition knowledge, interests, and information sources of male athletes. *J Nutr Educ* **18**, 243–245.
83. Hays R, Anderson R & Revicki D (1993) Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* **2**, 441–449.