Journal of
Clinical Nursing

ISSUES IN CLINICAL NURSING

Essential elements of questionnaire design and development

Janice Rattray PhD, MN, DipN, Cert Ed RGN, SCM
Senior Lecturer in Nursing, Postgraduate Student Advisor, School of Nursing and Midwifery, University of Dundee, Dundee, UK

Martyn C Jones PhD, C Psychol, RNMH Dip Ed, Dip NBS, ILTM
Senior Lecturer in Nursing, School of Nursing and Midwifery, University of Dundee, Dundee, UK

Submitted for publication: 7 April 2005
Accepted for publication: 20 April 2005

Correspondence:
Janice Rattray
School of Nursing and Midwifery
University of Dundee, Ninewells Hospital,
Dundee, DD1 9SY
Telephone: +44(0)1382 632304
E-mail: j.z.rattray@dundee.ac.uk

RATTRAY J & JONES MC (2007) *Journal of Clinical Nursing* **16**, 234–243

Essential elements of questionnaire design and development

Aims. The aims of this paper were (1) to raise awareness of the issues in questionnaire development and subsequent psychometric evaluation, and (2) to provide strategies to enable nurse researchers to design and develop their own measure and evaluate the quality of existing nursing measures.

Background. The number of questionnaires developed by nurses has increased in recent years. While the rigour applied to the questionnaire development process may be improving, we know that nurses are still not generally adept at the psychometric evaluation of new measures. This paper explores the process by which a reliable and valid questionnaire can be developed.

Methods. We critically evaluate the theoretical and methodological issues associated with questionnaire design and development and present a series of heuristic decision-making strategies at each stage of such development. The range of available scales is presented and we discuss strategies to enable item generation and development. The importance of stating *a priori* the number of factors expected in a prototypic measure is emphasized. Issues of reliability and validity are explored using item analysis and exploratory factor analysis and illustrated using examples from recent nursing research literature.

Conclusion. Questionnaire design and development must be supported by a logical, systematic and structured approach. To aid this process we present a framework that supports this and suggest strategies to demonstrate the reliability and validity of the new and developing measure.

Relevance to clinical practice. In developing the evidence base of nursing practice using this method of data collection, it is vital that questionnaire design incorporates preplanned methods to establish reliability and validity. Failure to develop a questionnaire sufficiently may lead to difficulty interpreting results, and this may impact upon clinical or educational practice. This paper presents a critical evaluation of the questionnaire design and development process and demonstrates good practice at each stage of this process.

Key words: nurses, nursing, psychometric evaluation, questionnaire design, scale construction

doi: 10.1111/j.1365-2702.2006.01573.x

Introduction

The use of questionnaires as a method of data collection in health-care research both nationally and internationally has increased in recent years (Sitzia *et al.* 1997, Bakas & Champion 1999, Chen 1999, Jones & Johnston 1999, Jeffreys 2000, Waltz & Jenkins 2001, Siu 2002, Rattray *et al.* 2004). The increasing emphasis on evidence-based health care makes it even more important that nurses understand the theoretical issues associated with such methods. When interpreting results from questionnaires, the development process should be defined in sufficient detail and with sufficient rigour to enable a practitioner to make an informed decision about whether to implement findings. We use questionnaires to enable the collection of information in a standardized manner which, when gathered from a representative sample of a defined population, allows the inference of results to the wider population. This is important when we want to evaluate the effectiveness of care or treatment. While the rigour applied to the questionnaire development process may be improving, nurses are still neither generally adept nor confident at the psychometric evaluation of such measures (Jones & Johnston 1999). Central to the understanding of results derived from questionnaires are the issues of reliability and validity which underpin questionnaire development from item generation, the proposal of an *a priori* factor structure to subsequent psychometric analysis.

Whilst relevant texts may provide information about these issues, rarely is sufficient detail provided in a single source to guide the questionnaire development process. This paper provides a critical analysis of key methodological issues from item generation to planned psychometric evaluation and collates a series of heuristic decision-making strategies to assist practitioners to develop their own measure or evaluate the work of others. Two worked examples illustrate these strategies drawn from clinical practice and nurse education (Jones & Johnston 1999, Rattray *et al.* 2004). Issues of reliability and validity are explored using item analysis and exploratory factor analytic techniques.

What will the questionnaire measure?

Nurse researchers use questionnaires to measure knowledge, attitudes, emotion, cognition, intention or behaviour. This approach captures the self-reported observations of the individual and is commonly used to measure patient perceptions of many aspects of health care (see Table 1 for examples). When developing a questionnaire, items or questions are generated that require the respondent to respond to a series of questions or statements. Participant responses are then converted into numerical form and statistically analysed. These items must reliably operationalize the key concepts detailed within specific research questions and must, in turn, be relevant and acceptable to the target group. The main benefits of such a method of data collection are that questionnaires are usually relatively quick to complete, are relatively economical and are usually easy to analyse (Bowling 1997).

This approach to data generation is not without criticism. It assumes that the researcher and respondents share underlying assumptions about language and interpret statement wording in a similar manner. Closed questions which are commonly used may restrict the depth of participant response (Bowling 1997) and thus the quality of data collected may be diminished or incomplete. Questionnaire-based methods are, therefore, not the method of choice where little is known about a subject or topic area. In such an instance, qualitative methods may be more appropriate.

The range of scales available

There are a range of scales and response styles that may be used when developing a questionnaire. These produce different types or levels of data (see Table 1) and this will influence the analysis options. Therefore, when developing a new measure, it is important to be clear which scale and response format to use. Frequency scales may be used when it is important to establish how often a target behaviour or event has occurred, e.g. the Intensive Care Experience Questionnaire (Rattray *et al.* 2004). Thurstone scales are less common in nursing research. Such scales use empirical data derived from judges to ensure that attitudes or behaviours being measured are spaced along a continuum with equal weighting/spacing, e.g. Nottingham Health Profile (Hunt *et al.* 1985). Guttman scaling is a hierarchical scaling technique that ranks items such that individuals who agree with an item will also agree with items of a lower rank, e.g. Katz Index of Activities of Daily Living (Katz *et al.* 1963). Rasch scaling is a similar type of scale, e.g. De Jong Gierveld and Kamphuis (1985) and Kline (1993). Knowledge questionnaires may be helpful when evaluating the outcome of a patient education programme, e.g. Furze *et al.* (2001). They generally offer multiple choice or dichotomous yes/no response options.

Within research in nursing Likert-type or frequency scales are most commonly used. These scales use fixed choice response formats and are designed to measure attitudes or opinions (Bowling 1997, Burns & Grove 1997). These ordinal scales measure levels of agreement/disagreement. A Likert-type scale assumes that the strength/intensity of experience is linear, i.e. on a continuum from strongly agree

Table 1 Stages in questionnaire development: item generation and scale construction

Questionnaire development	Key issues	Examples of measures
What will the questionnaire measure?	Knowledge Attitude/beliefs/intention Cognition Emotion Behaviour	The York Angina Beliefs Questionnaire, (Furze *et al.* 2001) Operationalising the Theory of Planned Behaviour (Conner & Sparks 1995) Illness Perception Questionnaire (Weinman *et al.* 1996) Anxiety, depression (Spielberger *et al.* 1983, Goldberg & Williams 1988) Functional Limitations Profile, FLIP (Patrick & Peach 1989)
What types of scale can be used?	Frequency Thurstone Rasch Guttman Mokken Likert type Multiple choice	ICEQ, (Rattray *et al.* 2004) Nottingham Health Profile, (Hunt *et al.* 1985) Loneliness scale (De Jong Gierveld & Kamphuis 1985) FLIP (Patrick & Peach 1989) Edinburgh Feeding Evaluation in Dementia, (Watson & Deary 1996) SNSI, (Jones & Johnston 1999) The York Angina Beliefs Questionnaire, (Furze *et al.* 2001)
How do I generate items for my questionnaire?	Ensure relevance of items? Wording issues Which response format is best? Which types of question are possible? Free text options? Does your measure have subscales? Questionnaire layout	Check research questions, explore literature, experts, target population Follow established guidelines (Oppenheim 1992, Bowling 1997). Discard poor items. Consider and pilot response format (five-point, seven-point, visual analogue scale) In standardized measures most are closed, to allow combination of scores from large numbers of respondents. May have some open, free text responses. Construct items that represent each different hypothesized domain Carefully consider order of items

to strongly disagree, and makes the assumption that attitudes can be measured. Respondents may be offered a choice of five to seven or even nine precoded responses with the neutral point being neither agree nor disagree. There is no assumption made that equal intervals exist between the points on the scale; however, they can indicate the relative ordering of an individual's response to an item. While this is perhaps too simplistic, until an alternative model is developed, it is a relatively easy and appropriate method to use (Oppenheim 1992). Some controversy exists as to whether a neutral point should be offered. If this option is removed, this forces the respondent to choose a response, which may lead to respondent irritation and increase non-response bias (Burns & Grove 1997).

It is acceptable to treat scores from this type of response format as interval data to allow the use of common parametric tests (Ferguson & Cox 1993, Polgar & Thomas 1995, Bowling 1997, Burns & Grove 1997). As with any data set, subsequent statistical analysis should be determined by the normality of distribution of the data and whether the data meets the underlying assumptions of the proposed statistical test.

It would be unusual to develop a questionnaire that relied upon a single-item response, and multi-item scales are generally used in preference to single-item scales to avoid bias, misinterpretation and reduce measurement error (Bowling 1997, Burns & Grove 1997). Such questionnaires have a number of subscales that 'tap' into the main construct being

measured. For example, the Short-Form 36 (Ware & Sherbourne 1992) measures health-related quality of life using 36 items representing eight health subscales.

Item generation, wording and order

The generation of items during questionnaire development requires considerable pilot work to refine wording and content. To assure face or content validity, items can be generated from a number of sources including consultation with experts in the field, proposed respondents and review of associated literature (Priest *et al.* 1995, Bowling 1997; see Table 1). In addition, a key strategy in item generation is to revisit the research questions frequently and to ensure that items reflect these and remain relevant (Oppenheim 1992, Bowling 1997). It is during this stage that the proposed subscales of a questionnaire are identified (Ferguson & Cox 1993) and to ensure that items are representative of these. The item and factor analysis stages of the questionnaire development process may then be used to establish if such items are indeed representative of the expected subscale or factor.

The type of question, language used and order of items may all bias response. Consideration should be given to the order in which items are presented, e.g. it is best to avoid presenting controversial or emotive items at the beginning of the questionnaire. To engage participants and prevent boredom, demographic and/or clinical data may be presented at the end. Certain questions should be avoided, e.g. those that lead or include double negatives or double-barreled questions (Bowling 1997). A mixture of both positively and negatively worded items may minimize the danger of acquiescent response bias, i.e. the tendency for respondents to agree with a statement, or respond in the same way to items.

To allow respondents to expand upon answers and provide more in-depth responses, free text response or open questions may be included. Respondents may welcome this opportunity. However, whilst this approach can provide the interviewer with rich data, such material can be difficult to analyse and interpret (Polgar & Thomas 1995). However, these problems may be outweighed by the benefits of including this option and can be especially useful in the early development of a questionnaire. Free text comments can inform future questionnaire development by identifying poorly constructed items or new items for future inclusion.

Piloting a questionnaire using item analysis ($N \leq 100$)

It is important to ensure that sufficient pilot work is carried out during the development of a new measure. This will identify items that lack clarity or that may not be appropriate for, or discriminate between, respondents. Ideally, the questionnaire should be piloted on a smaller sample of intended respondents, but with a sample size sufficient to perform systematic appraisal of its performance. Item analysis is one way to pilot a questionnaire. This provides a range of simple heuristics on item retention or deletion, see Table 2. High endorsement of an option within a particular item suggests poor discriminatory power or the redundancy of an item that requires deletion (Priest *et al.* 1995). Alternatively, a Cronbach's $\alpha < 0{\cdot}70$ may suggest that items in a questionnaire or subscale are poorly grouped. To identify specific items that do not add to the explanatory power of the questionnaire or subscale an item-total correlation cut-off of $< 0{\cdot}3$ can be used (Ferketich 1991, Kline 1993). However, it is important when revising the questionnaire to refer constantly to the original research questions that are being addressed and retain items that are thought to reflect the underlying theoretical domains of the questionnaire despite poor psychometric analysis. Problem items may also be identified because of high levels of non-response.

Demonstrating reliability

It is essential that the reliability of a developing questionnaire can be demonstrated. Reliability refers to the repeatability, stability or internal consistency of a questionnaire (Jack & Clarke 1998). One of the most common ways to demonstrate this uses the Cronbach's α statistic. This statistic uses inter-item correlations to determine whether constituent items are measuring the same domain (Bowling 1997, Bryman & Cramer 1997, Jack & Clarke 1998). If the items show good internal consistency, Cronbach's α should exceed 0·70 for a developing questionnaire or 0·80 for a more established questionnaire (Bowling 1997, Bryman & Cramer 1997). It is usual to report the Cronbach's α statistic for the separate domains within a questionnaire rather for the entire questionnaire.

Item-total correlations can also be used to assess internal consistency. If the items are measuring the same underlying concept then each item should correlate with the total score from the questionnaire or domain (Priest *et al.* 1995). This score can be biased, especially in small sample sizes, as the item itself is included in the total score (Kline 1993). Therefore, to reduce this bias, a corrected item-total correlation should be calculated. This removes the score from the item from the total score from the questionnaire or domain (Bowling 1997) prior to the correlation. Kline (1993) recommends deleting any questionnaire item with a corrected item-total correlation of $< 0{\cdot}3$. Item analysis using inter-item

Table 2 Stages in questionnaire development: piloting the questionnaire: item analysis (Adapted from Rattray *et al.* 2004)

Questionnaire development	Key issues	Examples of decision aids
Piloting the questionnaire: Item analysis	Spread of responses across options:	High endorsement of a single option is problematic (Priest *et al.* 1995). An item should be considered for removal if ≥80%, ≤20% of responses endorsed one response.
	Initial psychometric analysis:	Items with an inter-item correlation of $<0{\cdot}3$ or $>0{\cdot}7$ should be considered for removal (Ferketich 1991). Items with a poor Cronbach's α, i.e. $<0{\cdot}7$ should be considered for removal (Kline 1993).
	Clarity and relevance of items:	Researcher's interpretation of patient comments. Alternatively, if respondents fail to complete an item it suggests that the item may lack clarity.
	Items deemed theoretically important:	Items should be retained if they are deemed to be theoretically important even if they do not meet the above criteria.
	Is your measure affected by social desirability bias?	Explore the relationship between item and scale total with measure that captures this response tendency, e.g. Marlowe–Crown Social Desirability Index (Crowne & Marlowe 1960)
Reliability	Internal consistency	Corrected inter-item correlations (Ferketich 1991)
	Test–retest	Item-total correlation (Ferketich 1991)
	Inter-observer	Cronbach alpha (Kline 1993)
		Temporal stability of the measure (Johnson 2001)
		Observational studies (e.g. Ager 1998, Ager *et al.* 2001)
Validity	Face or content	Do the items sufficiently represent different hypothesized domains?
	Concurrent or discriminant	Do subscale scores correlate with existing, validated measures presented concurrently?
	Predictive	Do subscale scores predict hypothesis reports on existing, validated measures presented longitudinally?

correlations will also identify those items that are too similar. High inter-item correlations ($>0{\cdot}8$) suggest that these are indeed repetitions of each other (sometimes referred to as bloated specifics) and are in essence asking the same question (Ferketich 1991, Kline 1993).

Test–retest reliability can assess stability of a measure over time and this should be included in the process of any questionnaire development. This is of particular importance if the intended use of the measure is to assess change over time or responsiveness.

Demonstrating validity

Validity refers to whether a questionnaire is measuring what it purports to (Bryman & Cramer 1997). While this can be difficult to establish, demonstrating the validity of a developing measure is vital. There are several different types of validity (Polgar & Thomas 1995, Bowling 1997, Bryman & Cramer 1997). Content validity (or face validity) refers to expert opinion concerning whether the scale items represent the proposed domains or concepts the questionnaire is intended to measure. This is an initial step in establishing validity, but is not sufficient by itself. Convergent (or concurrent) and discriminant validity must also demonstrated by correlating the measure with related and/or dissimilar measures (Bowling 1997). When developing a questionnaire it is, therefore, important to include, within the research design, additional established measures with proven validity against which to test the developing questionnaire. Construct validity relates to how well the items in the questionnaire represent the underlying conceptual structure. Factor analysis is one statistical technique that can be used to determine the constructs or domains within the developing measure. This approach can, therefore, contribute to establishing construct validity.

Further development: exploratory factor analysis ($N > 100$)

Following initial pilot work and item deletion, the questionnaire should be administered to a sample of sufficient size to

allow exploratory factor analytic techniques to be performed. Ferguson and Cox (1993) suggest that 100 respondents is the absolute minimum number to be able to undertake this analysis. However, others would suggest that this is insufficient and a rule of thumb would be five respondents per item (Bryman & Cramer 1997). This type of analysis must follow a predefined and systematic analytic sequence (Ferguson & Cox 1993).

Principal components analysis (PCA) explores the interrelationship of variables. It provides a basis for the removal of redundant or unnecessary items in a developing measure (Anthony 1999) and can identify the associated underlying concepts, domains or subscales of a questionnaire (Oppenheim 1992, Ferguson & Cox 1993). The terms of factor analysis and PCA are often used synonymously in this context. In practice, however, PCA is most commonly used. Rarely is a questionnaire uni-dimensional and PCA usually identifies the presence of one principal component that accounts for most of the variance and subsequent components that account for less and less.

In the initial PCA analysis of an unrotated solution, most items should 'load', i.e. correlate with the first component. This can make interpretation of results difficult (Kline 1994), and to assist the interpretation of a factor solution, rotation of factors (components) is often performed. This should be a standard option on statistical packages, e.g. Statistical Package for Social Scientists (SPSS Inc., Chicago, IL, USA). Factor rotation maximizes the loadings of variables with a strong association with a factor, and minimizes those with a weaker one (Oppenheim 1992) and often helps make sense of the proposed factor structure. Varimax rotation, which is an orthogonal rotation (i.e. one in which the factors do not correlate), is often used, particularly if the proposed factors are thought to be independent of each other (Ferguson & Cox 1993). However, oblimin rotation may be used, when factors are thought to have some relationship, e.g. Jones and Johnston (1999). It is, therefore, vital to state *a priori* the number of factors you expect to emerge and to have decided which rotation method you will use ahead of any analysis.

Pre-analysis checks

Ferguson and Cox (1993) give a detailed account of the process of exploratory factor analysis and provide a set of heuristics for its three stages of pre-analysis checks, extraction and rotation (see Table 3 for the pre-analysis checks). These pre-analysis checks are necessary to ensure the proposed data set is appropriate for the method. The checks include determining the stability of the emerging factor structure, sampling requirements, item scaling, skewness and kurtosis of variables and the appropriateness of the correlation matrix.

Factor extraction

Two main methods are used to decide upon the number of emerging factors, Kaiser's criterion for those factors with an eigenvalue of >1 and the scree test. An eigenvalue is an estimate of variance explained by a factor in a data set (Ferguson & Cox 1993), and a value >1 indicates greater than average variance. A scree test is the graphic representation of this. Figure 1 shows the scree test that demonstrated the four-factor structure from the SNSI (Jones & Johnston 1999). The number of factors is identified from the break in the slope. If a straight line is fitted along the eigenvalue rubble, the number of domains within the questionnaire is revealed by the number of factors above the line. This latter method includes a degree of subjectivity in its interpretation.

With PCA, the removal of redundant items within a developing measure occurs within an iterative process. Agius *et al.* (1996) describe an iterative process of removing variables with general loadings (of 0·40 on more than one factor) and weak loadings (failing to load above 0·39 on any factor). This process is applied to the initial unrotated PCA before applying a varimax or oblimin rotation to interpret the structure of the solution. In the development of the SNSI, first unrotated principal component revealed the loading of 41 items accounting for 24·9% of the variance in the correlation matrix. The scree plot suggested a four-factor solution for rotation. Four further iterations of this variable reduction process led to the final 22-item solution, accounting for 51·3% of the variance in the correlation matrix.

Two recent examples of questionnaire development are the Intensive Care Experience Questionnaire (Rattray *et al.* 2004) and the Student Nurse Stress Index (Jones & Johnston 1999) for use in clinical and educational contexts respectively. Both measures used the questionnaire development approach described in this paper (see Table 4). In particular, the suitability of the data set for this type of analysis was established following the range of pre-analysis checks in each case. The questionnaires were piloted using both item and exploratory factor analysis. The hypothesized factor structure was demonstrated in the ICEQ (Rattray *et al.* 2004) but not in the SNSI (Jones & Johnston 1999) in which a fourth factor emerged. This finding demonstrates the exploratory nature of this type of factor analytic technique and the need to confirm findings in an independent data set (Agius *et al.* 1996). Confirmation of the initial four-factor structure was achieved in an independent data set with the SNSI. Work is currently being undertaken for the ICEQ. Both questionnaires

Table 3 Stages in questionnaire development: factor analysis

Questionnaire development	Key issues	Pre-analysis checks (Ferguson & Cox 1993)
Further development: Exploratory Factor analysis	Principal components analysis (PCA): Explores the inter-relationship of variables Provides a basis for the removal of redundant or unnecessary items (Anthony 1999), PCA is used to identify the underlying domains or factors within a measure. Prior to analysis, must propose an underlying theoretical structure Ensure that the data set is appropriate Must follow a predefined and systematic analytic sequence, e.g. Ferguson and Cox (1993)	Stable Factor Structure Minimum number of participants: 100 Minimum participant to variable ratio, *N*/*p*: 2:1–10:1 Minimum variable to factor ratio, *p*/*m*: 2:1–6:1 Minimum participant to factor ratio, *N*/*m*: 2:1–6:1 Sampling Random sampling from a population. Item scaling Likert, Mokken and frequency scales are acceptable. Normality of distribution/skewness and kurtosis Underlying assumption is of normal distribution. Values of skewness and kurtosis should be calculated for each variable, and values out with accepted levels dealt with appropriately. Appropriateness of the correlation matrix Kaiser Meyer–Olkin: can the correlations between variables be accounted for by a smaller set of factors? should be $>0{\cdot}5$. Bartlett Test of Sphericity: based on the chi-squared test, – a large and significant test used to indicate discoverable relationships
Further development: Confirmatory factor analysis	Allows the further testing of the construct validity of the measure	Confirmation of factor structure on an independent data set, using exploratory and confirmatory methods, see Agius *et al.* (1996), Jones and Johnston (1999). Same underlying assumptions as exploratory methods. Confirmatory process uses single sample and multi-sample approaches

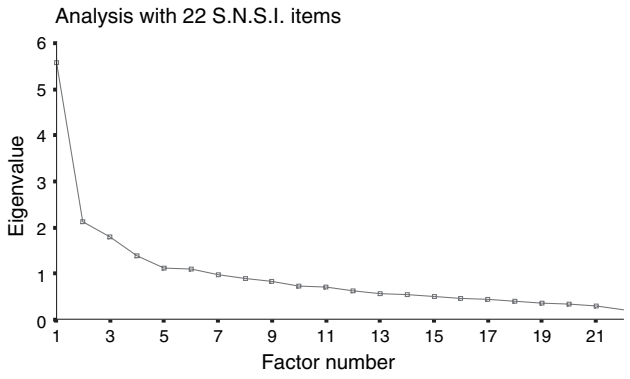

Figure 1 Scree test from the SNSI (Jones & Johnston 1999).

demonstrated good reliability and concurrent validity. For further details of the domain structure of the ICEQ and SNSI, see original papers (Jones & Johnston 1999, Rattray *et al.* 2004). These papers provide a step-by-step account of the questionnaire development process in a level of detail that is not available in traditional textbooks. This will be of particular use to the nurse researcher or research-minded practitioner.

Conclusion

This paper emphasizes the need to adopt a logical, systematic and structured approach to questionnaire development. We have presented a framework that supports this type of approach and have illustrated the questionnaire development process using item analysis, factor analytic and related methods and have demonstrated strategies to demonstrate the reliability and validity of the new and developing measure. We have suggested the need to preplan each stage of the questionnaire development process and provide a series

Table 4 Development of the ICEQ (Rattray *et al.* 2004) and SNSI (Jones & Johnston 1999)

	ICEQ (Rattray *et al.* 2004)	SNSI (Jones & Johnston 1999)
Purpose	The rationale for this questionnaire was identified from literature. Patients had limited recall of the ICU experience, yet described it as being frightening and persecutory in nature. Reported perceptions of this experience have been linked to poorer emotional outcome. Previous research in this field was mainly qualitative and, therefore, a standardized questionnaire was developed	The main purpose of this measure was to develop a reliable and valid questionnaire to measure the sources of stress for student nurses. Previous research had demonstrated high levels of distress associated with training to be a student nurse (Jones & Johnston 1997). It was important to identify the sources of stress for students, to inform a stress management intervention (Jones & Johnston 2000)
Research questions	Research questions were identified	A four-factor structure was hypothesized including academic load, clinical concerns and interface worries
Scale and response format	Likert-type and frequency scales with a five-choice format. Three open questions included	Likert-type items with a five-choice format
Generation of items	Items generated from experts, literature review and an underlying theoretical structure of five domains was proposed. Thirty-eight items generated, randomly placed throughout the measure, with a mix of positively and negatively worded items	An existing questionnaire with 43 items (Beck & Srivastava 1991). Fifteen additional items were generated from literature review and student feedback
Test and pilot of items	Pilot work: 34 patients interviewed	Pilot work was with a large data set of 320 students
Amendments based on item analysis or related techniques	Amendments made using criteria presented in Table 1. Eighteen items were removed, 11 were added leaving a 31-item questionnaire. Research questions revisited. Again the underlying theoretical structure of four domains was proposed	Item reduction carried out using exploratory factor analysis methods, rather than item analysis. Unrotated PCA. Weak Items (failing to load above 0·39) and general items (loading at or above 0·40 on more than one factor in the unrotated solution) were deleted in an iterative process
Principal component's analysis	Administered to 109 patients as part of a structured interview. Pre-analysis check ensured data were appropriate. Unrotated PCA Varimax rotation Factors with a loading of ≥0·4 on one factor only were retained. Items were reduced from 31 to 24. Four domains were identified	Forty-three plus 15 items were administered to 320 students. Pre-analysis check ensured data were appropriate. Oblimin rotation Items were reduced to a 22 item simple oblique solution. Four subscales were identified, academic load, clinical concerns, interface worries and personal problems
Reliability	Cronbach α statistic for each domain was ≥0·7	Cronbach α statistic for each domain was ≥0·73 (interface worries in an initial data set α 0·68)
Validity	Concurrent validity established by correlating domain scores with scores from two measures with demonstrated validity, e.g. Hospital Anxiety and Depression Scale, Impact of Event Scale	Concurrent validity was shown by correlating SNSI subscale scores with GHQ 30 (continuously scored). Discrminant validity demonstrated with distressed students scoring higher on all SNSI subscales
Confirmation on an independent data set	Data is being gathered to confirm the four-factor structure of the ICEQ	Four-factor structure was confirmed on an independent data set ($N = 195$) using exploratory and confirmatory factor analytic techniques (Deary *et al.* 1993)
Revision of measure		A revised 49-item version of the SNSI is currently in development (Jones & Johnston 2003)

of heuristic strategies to enable the nurse researcher to achieve this (Deary *et al.* 1993, Kline 1993, Agius *et al.* 1996).

While there has been an increase in the use of questionnaires within the nursing literature, few such measures have been developed using the full set of strategies used by Rattray *et al.* (2004) and Jones and Johnston (1999), summarized here. In developing the evidence base of nursing practice using this method of data collection, it is vital that the nurse researcher incorporates methods to establish the reliability and validity, particularly of new questionnaires. Failure to develop a questionnaire sufficiently may lead to difficulty interpreting results. For example, failure to demonstrate an expected correlation of a new measure with an established scale may arise because of limited variation in scores on a developing questionnaire and the subsequent suppression of correlations between scores on the two questionnaires. Alternatively, there may really be no reliable relationship between such variables. If a measure is poorly designed and has had insufficient psychometric evaluation, it may be difficult to judge between such competing explanations. In addition, it may not be possible to use the findings from an established measure, if that measure cannot be shown to be reliable in a particular sample.

If clinical or educational practice is to be enhanced or changed using findings derived from questionnaire-based methods, it is vital that the questionnaire has been sufficiently developed. This paper presents a critical evaluation of the questionnaire design and development process and demonstrates good practice at each stage of this process. This paper will enable the informed nurse researcher to plan the design and development of their own questionnaire, to evaluate the quality of existing nursing measures, and to inspire confidence in applying findings into practice.

Contributions

Study design: JR, MCJ; data analysis: JR, MCJ and manuscript preparation: JR, MCJ.

References

Ager A (1998) *The British Institute of Learning Disabilities Life Experiences Checklist*. BILD Publications, Kidderminster.

Ager A, Myers F, Kerr P, Myles S & Green A (2001) Moving home: social integration for adults with intellectual disabilities resettling into the community. *Journal of Applied Intellectual Disabilities* **14**, 392–400.

Agius RM, Blenkin H, Deary IJ, Zealley HE & Wood RA (1996) Survey of perceived stress and work demands of consultant doctors. *Occupational and Environmental Medicine* **53**, 217–224.

Anthony D (1999) *Understanding Advanced Statistics: A Guide for Nurses and Health Care Researchers*. Churchill Livingstone, Edinburgh.

Bakas T & Champion V (1999) Development and psychometric testing of the Bakas caregiving outcomes scale. *Nursing Research* **48**, 250–259.

Beck DL & Srivastava R (1991) Perceived level and sources of stress in baccalaureate nursing students. *Journal of Nursing Education* **30**, 127–133.

Bowling A (1997) *Research Methods in Health*. Open University Press, Buckingham.

Bryman A & Cramer D (1997) *Quantitative Data Analysis with SPSS for Windows*. Routledge, London.

Burns N & Grove SK (1997) *The Practice of Nursing Research Conduct, Critique, & Utilization*. W.B. Saunders and Co., Philadelphia.

Chen M-L (1999) Validation of the structure of the perceived meanings of cancer pain inventory. *Journal of Advanced Nursing* **30**, 344–351.

Conner M & Sparks P (1995) The theory of planned behaviour and health behaviours. In *Predicting Health Behaviour* (Conner M & Norman P eds). Open University Press, Buckingham, pp. 121–162.

Crowne DP & Marlowe DA (1960) A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology* **24**, 349–354.

De Jong Gierveld J & Kamphuis F (1985) The development of a Rasch-type loneliness scale. *Applied Psychological Measurement* **9**, 289–299.

Deary IJ, Hepburn DA, MacLeod KM & Frier BM (1993) Partitioning the symptoms of hypoglycaemia using multi-sample confirmatory factor analysis. *Diabetologia* **36**, 771–777.

Ferguson E & Cox T (1993) Exploratory factor analysis: a user's guide. *International Journal of Selection and Assessment* **1**, 84–94.

Ferketich S (1991) Focus on psychometrics: aspects of item analysis. *Research in Nursing and Health* **14**, 165–168.

Furze G, Lewin RJP, Roebuck A, Thompson DR & Bull P (2001) Attributions and misconceptions in angina: an exploratory study. *Journal of Health Psychology* **6**, 501–510.

Goldberg D & Williams P (1988) *A Users Guide to the General Health Questionnaire*. NFer-Nelson, Windsor.

Hunt S, McEwen J & McKenna SP (1985) Measuring health status: a new tool for clinicians and epidemiologists. *Journal of the Royal College of General Practitioners* **35**, 185–188.

Jack B & Clarke A (1998) The purpose and use of questionnaires in research. *Professional Nurse* **14**, 176–179.

Jeffreys MR (2000) Development and psychometric evaluation of the transcultural self-efficacy tool: a synthesis of findings. *Journal of Transcultural Nursing* **11**, 127–136.

Johnson J (2001) Evaluation of learning according to objectives tool. In *Measurement of Nursing Outcomes* (Waltz C & Jenkins L eds). Springer Publishing Company, New York, pp. 216–223.

Jones MC & Johnston DW (1997) Distress, stress and coping in first-year student nurses. *Journal of Advanced Nursing* **26**, 475–482.

Jones MC & Johnston DW (1999) The derivation of a brief student nurse stress index. *Work and Stress* **13**, 162–181.

Jones MC & Johnston DW (2000) Evaluating the impact of a worksite stress management programme for distressed student nurses: a randomised controlled trial. *Psychology and Health* **15**, 689–706.

Jones MC & Johnston DW (2003) *Further Development of the SNSI*. Paper presented at the Royal College of Nursing Annual International Research Conference, University of Manchester, April 2003.

Katz S, Ford A & Moskowitz R (1963) Studies of illness in the aged: the index of ADL. A standardised measure of biological and psychosocial function. *Journal of American Medical Association* **185**, 914–919.

Kline P (1993) *The Handbook of Psychological Testing*. Routledge, London.

Kline P (1994) *An Easy Guide to Factor Analysis*. Routledge, London.

Oppenheim AN (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. Pinter, London.

Patrick D & Peach H (eds) (1989) *Disablement in the Community*. Oxford University Press, Oxford.

Polgar S & Thomas S (1995) *Introduction to Research in the Health Sciences*. Churchill Livingstone, Melbourne.

Priest J, McColl BA, Thomas L & Bond S (1995) Developing and refining a new measurement tool. *Nurse Researcher* **2**, 69–81.

Rattray JE, Johnston M & Wildsmith JAW (2004) The intensive care experience: development of the intensive care experience (ICE) questionnaire. *Journal of Advanced Nursing* **47**, 64–73.

Sitzia J, Dikken C & Hughes J (1997) Psychometric evaluation of a questionnaire to document side-effects of chemotherapy. *Journal of Advanced Nursing* **25**, 999–1007.

Siu O-L (2002) Predictors of job satisfaction and absenteeism in two samples of Hong Kong Nurses. *Journal of Advanced Nursing* **40**, 218–229.

Spielberger C, Gorsuch R & Lushene R (1983) *The State-Trait Inventory: Test Manual for Form Y*. Consulting Psychologists Press, Palo Alto, CA.

Waltz C & Jenkins L (2001) *Measurement of Nursing Outcomes: Volume 1: Measuring Nursing Performance in Practice, Education, and Research*. Springer Publishing Company, New York.

Ware JE & Sherbourne CD (1992) The MOS 36-Item short-form health survey (SF-36): conceptual framework and item selection. *Medical Care* **30**, 473–481.

Watson R & Deary I (1996) Is there a relationship between feeding difficulty and nursing interventions in elderly people with dementia? *Nursing Times Research* **1**, 44–54.

Weinman J, Petrie K, Moss-Morris R & Horne R (1996) The illness perception questionnaire: a new measure for assessing the cognitive representation of illness. *Psychology and Health* **11**, 431–445.