Yu J & Kirk M (2009) Evaluation of empathy measurement tools in nursing: systematic review. *Journal of Advanced Nursing,* 65(9), 1790-1806.

Contact:
Dr Juping Yu, Faculty of Life Sciences and Education, University of South Wales, Glyntaf, Pontypridd, UK
juping.yu@southwales.ac.uk

Evaluation of empathy measurement tools in nursing: systematic review

ABSTRACT

Aim
This paper is a report of a systematic review conducted to analyse, evaluate and synthesize the rigour of measures used in nursing research to assess empathy, in order to identify a 'gold standard' for application in future studies.

Background
Empathy is considered essential to the provision of quality care. We identified 20 different empathy measures used in nursing research. There are inconsistencies between tools, indicating both the inherent complexity of measuring empathy and the need to evaluate the rigour of the measures themselves.

Data sources
An extensive search was conducted for the period 1987 and 2007 using the Medline, CINAHL and PsycINFO databases and the keywords 'empathy', 'tool', 'scale', 'measure', 'nurse' and 'nursing'. Twenty-nine studies were identified as relevant, in which 20 different empathy measurement tools were used. Twelve tools met the inclusion criteria for this review.

Method
Twelve measures were critically reviewed and analysed. A 7-criterion framework was developed appraising the rigour of empathy measures, with a range of 0-14 for each measure.

Results
Quality scores obtained were low (2-8 out of 14). Validity and reliability of data were commonly reported, but responsiveness to change was tested in only three measures. None of the measures were psychometrically robust or covered all the domains of empathy. User involvement was limited and only five were developed in nursing settings.

Conclusion
Most measures have undergone rigorous development and psychometric testing, although none is both psychometrically and conceptually satisfactory. Empathy measures need to cover all relevant domains reflecting users' own perspectives and be tested with appropriate populations in relevant care settings.

SUMMARY

What is already known about this topic

- Empathy is an essential component of the nurse-patient relationship and is crucial to quality nursing care.
- Twenty different measures to assess empathy in nursing research have been reported in the last 20 years.
- There are inconsistencies between the tools measuring empathy in nursing research, indicating the need for an evaluation of their rigour.

What this paper adds

- A framework for a consistent approach to assess the rigour of empathy measures applied in nursing research.
- None of the 12 measures of empathy reviewed is both psychometrically and conceptually satisfactory.
- Empathy measures need to reflect users' own perspectives and be tested in relevant settings.

Implications for practice and/or policy

- Tools used in the measurement of empathy in nursing research should be developed in the nursing context.
- A tool should give insight into users' views by involving them in the development of initial items and during the validation process.
- Evidence is needed that a tool can be used practically to assess empathy either through self-assessment or by patient-rating in clinical settings.

KEY WORDS

Empathy; measurement; nursing; patient care; quality of care; rigour; systematic review

INTRODUCTION

In his keynote speech to the United Kingdom National Health Service Confederation, Member of Parliament and Health Secretary Alan Johnson stated that compassionate care is crucial to the recovery of patients and that nursing quality should be measured according to the levels of care and empathy nurses exhibit with patients (Department of Health 2008). This reflects the wide recognition that empathy is a fundamental component of the nurse-patient relationship and of quality nursing care (Reynolds *et al.* 1999; Alligood 2005). However, if levels of empathy demonstrated by nurses are to be used as a measure of quality care, it is essential that the measurement tools applied are robust and methodologically rigorous.

Rogers (1957, p.99) defined empathy as the ability "to sense the client's private world as if it were your own, but without ever losing the 'as if' quality". Four key dimensions of empathy have been suggested: cognitive, emotive, moral and behavioural (Morse *et al.* 1992). The cognitive aspect shows the intellectual ability to identify and understand others' perspectives and predict their thoughts, the emotive dimension describes the ability to experience and share in others' psychological states or intrinsic feelings, the moral aspect refers to an internal altruistic drive that motivates the practice of empathy, and the behavioural dimension shows the ability to communicate empathetic understanding and concerns.

The reported empathy levels of nurses vary. Some researchers have found a high level of self-reported empathy (Bailey 1996, Watt-Watson *et al.* 2000), but a low level of empathy has been reported by others (Daniels *et al.* 1988, Reid-Ponte 1992). This may reflect the difficulties encountered in measuring empathy and the fact that the different measures were used to assess empathy in widely divergent populations. Higher empathy levels of nurses or nursing students have often been associated with positive patient outcomes, such as reduced distress and anxiety levels and increased likelihood of identifying the perceived needs of patients and carers (Murphy *et al.* 1992, Reid-Ponte 1992, Olson 1995, Olson & Hanchett 1997). A null correlation between empathy and patient outcomes, such as satisfaction, pain intensity and analgesic admission, has also been reported (Warner 1992, Watt-Watson *et al.* 2000). These inconsistent results may reflect the inherent complexity in measuring empathy, which is subjective, multi-faced and intangible, but nonetheless the rigour of the measures themselves needs also to be questioned.

In our earlier paper on empathy measurement in nursing research, an initial electronic search yielded 557 articles, indicating the wide interest, complexity and importance of this area (Yu & Kirk 2008). Twenty distinct tools were identified from 29 studies included in this earlier paper. We did not evaluate the quality of measures themselves, but gave a comprehensive overview of the measurement of empathy in nursing research. This showed that it is still unclear whether current empathy measurement tools are psychometrically and conceptually rigorous. This paper is sequential to our earlier paper.

THE REVIEW

Aim

The aim of the review was to analyse, evaluate and synthesize the rigour of measures used in nursing research to assess empathy, in order to identify a 'gold standard' for application in future studies.

Design

We conducted a two-phase methodological review. In phase 1, we reviewed studies reporting primary nursing research measuring empathy (Yu & Kirk 2008). The present paper focuses on the second phase, where we reviewed the tools themselves that were applied in these studies, with comparative analysis conducted against a quality framework. The Centre for Reviews and Dissemination (2001) guidelines on undertaking systematic reviews were followed, with reference to the evaluative checklist of Greenhalgh *et al.* (1998) for reviewing outcome measures.

Search Methods

Initially, literature searches were performed in CINAHL, MEDLINE and PsycINFO databases, using the keywords ‘empathy’, ‘tool’, ‘scale’, ‘measure’, ‘nurse’ and ‘nursing’, either alone or in combination, to identify relevant literature published between 1987 and 2007. Twenty-nine papers, using 20 different tools, reported primary nursing research measuring empathy (Yu & Kirk 2008). In this second phase, we focused on the 20 tools previously identified. The following criteria were used.

Inclusion criteria

Tools were only included if they:

- Described the original development of empathy measures
- Reported some psychometric properties (e.g. reliability, validity or responsiveness)
- Were published in English.

Exclusion criteria

Tools were excluded if they:

- Did not report any psychometric properties
- Were unavailable at the local library, via electronic journals or through the inter-library loan service
- Were written in a language other than English.

Search outcome

The relevance of retrieved literature compared with the inclusion and exclusion criteria was assessed by JY and checked by MK. Disagreement was resolved by checking the full text of papers and through further discussion until final agreement was reached. A flow diagram of the search process is shown in Figure 1. The 12 measures that satisfied the inclusion criteria were included in this review.

Quality appraisal

There are no published quality criteria for evaluating the rigour of measurement tools, although criteria are available for assessment of other types of studies, such as experimental, observational and qualitative studies. A 7-criterion appraisal framework was therefore developed, with reference to the work of Greenhalgh *et al.* (1998), Russell *et al.* (1998) and Grange *et al.* (2007). The framework was applied to each measure and the total score possible for each measure ranged from 0 to 14 (Table 1). A score of two points was awarded where the criterion was met, one point was awarded where the criterion was partially met, and zero was awarded where the criterion was not met.

Psychometrically, a robust tool should be valid, reliable and responsive (Bowling 2002, Polit & Beck 2004). The first criterion is validity, referring to whether a tool measures what is intended to measure. Validity may be described as comprising three major factors: face/content validity, construct validity and criterion validity (Streiner & Norman 2003, Polit & Beck 2004). Face/content validity is often determined by experts to check whether an instrument consists of adequate items for the construct being measured. Construct validity is the ability of a tool to measure the underlying concept it purports to measure. This can be established by:

1. correlating with measures that assess the same construct
2. the differences between groups (known-groups technique)
3. correlating with conceptually relevant measures (convergent validity)
4. not correlating with conceptually irrelevant measures (discriminant validity).

There are two types of criterion validity: concurrent validity and predictive validity. The former refers to how each measure correlates with a valid pre-existing measure of the same concept, and the later is the ability of a measure to predict future changes in key variables in an expected direction.

The second criterion addresses reliability, or the ability of a tool to measure in a reproducible fashion. It involves three key aspects: internal consistency, stability and equivalence (Streiner & Norman 2003, Polit & Beck 2004). Internal consistency shows whether all items of a tool are related and is commonly supported by using the split-half technique, Cronbach's alpha and item-total correlations. Stability is an assessment of whether a measure produces the same result for the same individual on different occasions (test-retest reliability). Equivalence demonstrates whether different raters produce the same results when independently rating an individual (inter-rater reliability).

The third criterion is responsiveness. This is the ability of a measure to detect actual change correctly over a pre-specified period following an intervention, and to identify whether individuals could demonstrate change using a reference measure (Husted *et al.* 2000, Roach 2006). There are three key aspects to the measurement of change: differences between individuals in the amount of change, factors associated with a good outcome and treatment effects from group differences (Linn & Slinde 1977).

The fourth criterion considers the setting in which a measure is developed and the fifth examines the degree of user involvement. A conceptually sound tool needs to be user-centred. It should be developed and validated with a defined population within a specific setting and comprise appropriate domains and items that are most relevant to users' own views (Greenhalgh *et al.* 1998). The level of user-centredness may vary along a spectrum from taking no account of user perceptions, simply listening to their views, to actively encouraging them to specify which factors should be recorded and how they should be defined. The sixth criterion addresses the domains of a scale, which should reflect user views. Lastly, the seventh criterion of practicality and application suggests the need for a measure to be practical, feasible and easy to completion for both participants and administrators (Greenhalgh *et al.* 1998).

Data extraction

Data extraction was guided by the quality appraisal framework developed. The data extracted are presented in two tables. Table 2 gives general information about the measures, comprising of bibliography, origin, population, domains, items, administration and nursing studies using the measure. Table 3 shows the psychometric properties of the tools, in terms of validity, reliability and responsiveness.

RESULTS

In total, 12 empathy measures were included in the review, each comprised of between three and 84 items (Table 2). A general description and critique of each measure is reported first, followed by an evaluation of their rigour.

General description and critique of the measures

Barrett-Lennard Relationship Inventory – empathic understanding (BLRI)

This measure was developed to assess the therapist-client relationship (Barrett-Lennard 1962). The empathic understanding subscale was tested with a small number of clients and therapists. It had good content, construct and predictive validity and high levels of split-half and test-retest reliability. The mean subscale inter-correlations were not strong. This self-rating tool did not take clients' or therapists' views into consideration in developing the items and has not been widely used in nursing.

Carkhuff Indices of Discrimination and Communication (CIDC)

This measure, including discrimination and communication indices, was developed for use in a helping situation (Carkhuff 1969a). The discrimination index involved a prospective helper rating 64 responses, and there was some evidence of its construct validity (Carkhuff 1969b). The communication index contained 16 expressions of a client's feelings. It had good test-retest and inter-rater reliability. The indices have also been shown to be sensitive to change among a small number of teachers (n=8) attending a training programme (Carkhuff 1969c). Most scenarios in the indices were female-focused and user views were not taken into account in developing the scale.

Emotional Empathy Tendency Scale (EETS)

This 33-item measure was validated for undergraduate students attending a psychology course (Mehrabian & Epstein 1972). It included seven interrelated subscales (r=0.30, P=0.01) to assess the emotional aspect of empathy and recognition of and sharing of others' feelings. It showed weak construct validity, but satisfactory discriminant validity. Reliability assessed by split-half technique was satisfactory, but levels of subscale inter-correlations ranged from low to moderate. The measure was not developed in healthcare settings and user perceptions were not considered in its development.

Emotional Intelligence Scale (EIS)

This 33-item, 5-point scale was developed with the general population to assess emotional intelligence (Schutte *et al.* 1998). Psychometrics were tested comprehensively with a small sample of college students and the general population. The instrument showed low to moderate levels of construct and predictive validity, but high levels of internal consistency and test-retest reliability. User views were not reflected in generating of the scale items, it was not developed in healthcare settings and sensitivity to change has not been examined.

Empathy Construct Rating Scale (ECRS)

La Monica (1981) developed this 84-item scale with nurses and patients in hospital settings. Initially, a pool of 500 items was generated by female graduate students from psychology (n=25) and nursing (n=25). Its face and content validity were judged by experts and nurses. It exhibited high levels of internal consistency, split-half reliability and test-retest reliability. Discriminant validity was weak, convergent validity was not evident, and inter-rater reliability and responsiveness were not assessed. This comprehensive but lengthy tool was developed for nurses and has been widely employed in nursing research. However, patient perspectives were not taken into account in developing the scale.

Hogan Empathy Scale (HES)

This scale was developed with the general population to assess intellectual appreciation of the feelings of others (Hogan 1969). A raw score ranging from 0-39 was produced, with higher scores indicating greater empathic ability. Construct validity was supported by statistically significant group and gender differences in empathy levels. It showed low to moderate levels of concurrent validity and acceptable levels of internal consistency and test-retest reliability. The measure was not developed specifically for healthcare professionals and it is not able to assess empathic behaviour.

Interpersonal Reactivity Index (IRI)

Davis (1980) developed this measure with the general population to assess cognitive and emotional aspects of empathy. The 28-item scale consisted of four 7-item subscales: perspective-taking, fantasy, empathic concern and personal distress. Construct validity has been established by factor analysis and statistically significant gender differences, with females scoring higher levels than males. Subscale inter-correlations were low or statistically non-significant, but it showed moderate levels of internal consistency and test-retest reliability. The scale was not developed for healthcare professionals and user perspectives were not considered in the development of the items.

Jefferson Scale of Physician Empathy (JSPE)

This 20-item, 7-point scale was developed to measure medical students' attitudes towards physician empathy (Hojat *et al.* 2001). Face validity was judged by physicians (n=100), while construct validity was obtained by factor analysis and gender comparison. It showed acceptable levels of convergent and discriminant validity and high levels of internal consistency, but sensitivity to change was absent. A modified version (HP-version) has been developed for physicians and other healthcare professionals (Hojot *et al.* 2002). Kliszcz *et al.* (2006) adapted this scale to a Polish version. The scale was not developed in nursing settings and user perspectives were not considered in its item generation.

Layton Empathy Test (LET)

Layton (1979) developed this scale with nursing students. There were two forms (I & II) and each had three parts. Levels of construct validity and reliability coefficients were low. It also showed unsatisfactory levels of responsiveness in the case of a small number of nurses who attended a training programme. Junior students showed a treatment effect for the Form II, but senior students did not. The scale was validated for nursing students, but their views were not taken into account in its item development.

Perception of Empathy Inventory (PEI)

This 4-point scale comprising 33 true/false statements was developed with hospitalised patients (n=81) to assess their perceptions of nurses' empathy (Wheeler 1990). The initial items were generated from a literature review and items of Empathy Understanding scale (Barrett-Lemmard 1962). Face and content validity have been demonstrated by four patients and two professors in psychiatric nursing. It showed acceptable levels of construct validity and a high level of internal consistency, but item-to-total correlations ranged from as low as 0.21 to 0.83. This scale was developed for patients to use, but their views were not sought in its item generation.

Reynolds Empathy Scale (RES)

This 12 item, 7-point, rater-rating scale was developed with nurses in the UK (Reynolds 2000). Face and content validity were examined by six experts from nursing and clinical psychology. The tool had high levels of concurrent validity, internal consistency, discrimination and test-retest reliability. Inter-rater reliability was demonstrated by final agreement between raters, reaching from 41.6% to 91.6%. Its responsiveness was examined among nurses attending a training programme. The mean score of respondents (n=22) in the experimental group (M=48.05, SD=9.81) was higher than those (n=15) in the control group (M=23.61, SD=6.95). There was a statistically significant change between pre-course, inter-training (t=6.84, $P<0.01$) and post-course levels (t=6.01, $P<0.01$). Marco *et al.* (2004) adapted this scale to a Spanish version. Patients were not involved in assessment of their perception of nurses' empathy, although their views were considered in generating the scale items.

Visual Analogue Scale (VAS)

Wheeler *et al.* (1996) developed this scale to assess teachers' perceptions of nursing students' empathic ability. The scale has not been tested comprehensively. It showed a low level of concurrent validity and satisfactory internal consistency reliability, while evidence of other psychometric properties was absent. Users' views were also not considered in developing the scale.

Assessment against the seven criteria

Validity

Validity was addressed in some way for all the measures (Table 3) and nine showed more than one type of validity (BLRI, ECRS, EIS, HES, IRI, JSPE, LET, PEI, RES). Construct validity was the most frequently reported method, described for all but two scales (RES, VAS). Moderate to high validity was established by:

- factor analysis (ECRS, EIS, IRI, JSPE)
- gender differences (HES, EETS, EIS, IRI, JSPE)
- group differences (BLRI, CIDC, EIS, HES)
- correlations between empathy and other variables (EETS, PEI).

Construct validity was also confirmed by convergent validity (EIS, JSPE, LET) and discriminant validity (ECRS, EETS, EIS, JSPE).

Criterion validity was reported for five measures by testing for concurrent validity through comparing with a 'gold standard' empathy measure (HES, RES, VAS) or by predictive validity through assessing the ability to predict future changes (BLRI, EIS). Reported criterion validity was low, with the exception of the RES. Face and content validity were evaluated by a panel of experts (BLRI, ECRS, HES, IRI, JSPE, LET, RES, VAS) and patients (PEI).

Reliability

Reliability data were presented for all measures (Table 3). Internal consistency was the most frequently-used method, reported for all but one measure (CIDC). This was demonstrated by:

- a Cronbach's alpha coefficient (ECRS, EIS, IRI, JSPE, PEI, RES, VAS)
- subscale inter-correlations (BLRI, EETS, IRI)
- item-total correlations (PEI)
- internal discrimination (RES)
- split-half technique (BLRI, ECRS, EETS)
- Kuder-Richardson coefficients (HES, LET).

Most tools had a moderate to high level of internal reliability, with alphas ranging from 0.70 to 0.98 and split-half correlation coefficients of more than 0.84. Item-total correlation coefficients were low, ranging from 0.21 to 0.83 (PEI), while subscale correlations were low or not statistically significant (IRI).

Stability was addressed for six measures via test-retest reliability with two weeks to 75 days interval between testing (BLRI, CIDC, ECRS, EIS, HES, IRI, RES). Moderate to high reliability was shown, with correlation coefficients ranging from over 0.61 (IRI) to 0.98 (ECRS).

Equivalence was reported for two measures, demonstrated by inter-rater reliability (CIDC, RES). Carkhuff (1969b) showed good reliability (r=0.89) for the CIDC. The initial agreement between raters for the RES was low, but the final agreement reached 41.6-91.6% (Reynolds 2000). The ECRS and VAS involved third-party ratings, but evidence for inter-rater reliability was absent.

Responsiveness

An assessment of responsiveness was conducted for only three measures (CIDC, LET, RES). They each exhibited ability to detect change to some extent following training. Pre- and post-training differences were assessed for intervention and control groups through repeated measures analysis of variance (LET, RES) and Spearman rank-order correlations (CISC). Charkhuff (1969c) also tested the ability of the Communication and Discrimination Indexes to identify individual trainees with different amounts of change. Those entering training above level 1.7 gained statistically significantly more and functioned at statistically significantly higher absolute levels of functioning following training with a high-level trainer.

Setting

Of the 12 measures, 11 were originally developed in the USA, while the Reynolds (2000) Empathy Scale was the only one generated in the UK (Table 2). Seven were developed and tested in disciplines other than nursing, four being developed with the general population (EETS, EIS, HES, IRI), two in the counselling context (BLRI, CIDC) and one in a medical setting (JSPE). Of the five measures focusing on nursing, two were tested with nursing students (LET, VAS), one with nurses (RES), one with patients (PEI) and one with both nurses and patients (ECRS).

User-centeredness

Only one measure (PEI) involved patients (n=2) in the test for face and content validity. In two measures (ECRS, RES), user views were taken into account in initial item generation. An initial pool of items was normally generated from literature reviews with or without a guiding theory. These included Rogers' (1957) theory on client-centred therapy (BLRI, EETS, LET, PEI, RES), the model of emotional intelligence (Salovey & Mayer 1990) (EIS) and the theory of moral development (Hogan 1969) (HES).

Domain

Four domains were assessed: cognitive, emotional, moral and behavioural (Table 2). The behavioural domain was the most frequently assessed and was included in seven measures (BLRI, CIDC, ECRS, LET, PEI, RES, VAS). Two measures (ECRS, PEI) assessed patient-perceived empathic behaviour and four (CIDC, ECRS, RES, VAS) measured this behaviour as observed by others. The cognitive domain was assessed in six measures (CIDC, ECRS, HES, IRI, JSPE, LET). Four measures (EIS, ETI, HES, IRI) included the emotional dimension and one (HES) assessed the moral domain.

Practicality and application

Reported time taken for completion was available for one measure (LET), which took 10-15 minutes to complete. Six measures were self-administrated only (EETS, EIS, HES, IRI, JSPE, LET), four were third-party-ratings (CIDC, PEI, RES, VAS), one included both self and client ratings (BLRI) and one used self-, patient- and peer-ratings (ECRS). In three measures (CIDC, RES, VAS) training needs for rating were reported.

La Monica's (1981) ECRS scale is the most popular measure, and has been used in 10 nursing studies: in hospital (La Monica 1987, Murphy *et al.* 1992, Warner 1992, Bailey 1996), community (Astrom *et al.* 1990, 1991, Kuremyr *et al.* 1994, Palsson *et al.* 1996) and university settings (Daniels *et al.* 1988, Reynolds & Presly 1988). Seven measures (58%) have only been used in a single study (BLRI, CIDC, EETS, EIS, PEI, RES, VAS) and four (33%) were applied in three studies (HES, IRI, JSPE, LET).

Quality assessment

Quality scores raning from 0 to14 for each measure were calculated against the seven criteria (Table 4). The highest score was eight (ECRS) and the lowest was 2 (EETS). User-centeredness received the lowest score of only 2 for all the measures, while reliability had the highest score, 14 in total. Six measures (EETS, HES, IRI, JSPE, PEI, RES) partially met the validity and reliability criteria, scoring 1 for each. Three measures (CIDS, ECRS, EIS) scored 1 for validity, but 2 for reliability. One measure (BLRI) met validity criteria, scoring 2, but scored 1 for reliability. One scale (VAS) partially met reliability criteria and another (LET) satisfied neither validity nor reliability criteria.

DISCUSSION

Review limitations

The psychometric, conceptual and practical characteristics of the 12 measures used in nursing research were evaluated. This review provides a reference for nurses and researchers seeking guidance on how to select quality measures for assessing empathy in the nursing context. There are two limitations to the review. First, the exclusion of non-English publications, may have led to omission of some relevant scales in use in the measurement of empathy in nursing research, although they would need to be re-evaluated if translated and applied in a different setting. Second, the method of scoring was subjective, and so caution is needed when interpreting the quality scores.

To overcome this limitation and to increase reliability, both authors discussed, carefully checked and agreed with the scoring. Based on this review, the rigour of empathy measurement tools for use in nursing can be easily identified and a number of recommendations can be made.

Rigour of the measures

Psychometric testing was often limited to small populations, ranging from as small as 21 (Barrett-Lennard 1962) or from 3 to 8 in each subgroup (Layton 1979). Thus, sufficient power cannot be guaranteed to conduct statistical analysis. Most measures partially satisfied the validity and reliability criteria. However, only three (CIDC, LET, RES) had data on responsiveness to interventions, a defining feature of an outcome measure (Greenhalgh *et al.* 1998). A measure needs to be able to detect change when it does occur. Without testing for sensitivity to change, a measure’s ability to evaluate an intervention accurately remains questionable.

User-centeredness, a fundamental aspect of measurement tools, was absent from most measures. A tool cannot accurately capture the content of patients’ perceptions and the ways their views are expressed without actively involving them in its development. Greenhalgh *et al.* (1998) argue that a measure should always be developed with a relevant population and include information that reflects user perspectives. Otherwise, a tool is deemed to be invalid regardless of its psychometrical rigour. Only five measures were tested in the nursing context. Of these, the Empathy Construct Rating Scale (La Monica 1981) has been the most popular and received the highest score on quality appraisal. However, this scale does not give insight into patient views. Data on responsiveness were also unavailable and the scale is lengthy to complete.
In addition, this tool was originally developed in the USA. Measures with good reliability in one country may not be as reliable in other countries, even where there is a common language (Williams *et al.* 2001).

Questions are raised as to how empathy is measured. Six measures solely used self-assessment (Table 2). These tools are suitable to measure the cognitive, moral and emotional aspects of empathy. However, self-reporting bias may occur and these measures cannot be applied to assess empathic behaviour. Three measures (CIDC, RES, VAS) were developed to assess empathic behaviour observed by a trained judge or a peer based on participants’ empathic performance. This method of measurement is more objective than self-assessment. However, it raises a question about the accuracy of interpretation of the behaviour being measured. Inter-rater reliability established in a tool’s original development cannot be assumed in other studies. In addition, non-verbal interactions and respondents’ attitudes cannot be captured.

Three scales (BLRI, ECRS, PEI) involve patient ratings. Only this type of assessment can evaluate patients’ appreciation of nurses’ empathic behaviour, as described in the final phase of Barrett-Lennard’s (1981) multi-dimensional empathy cycle (Figure 2). However, none of them give insight into patient perspectives. The Empathy Understanding subscale (Barrett-Lennard 1962) uses both self-rating and client-rating, while the Empathy Construct Rating Scale (La Monica 1981) involves self-rating, patient-rating and peer-rating. As discussed above, different aspects of empathy can be assessed by different methods. Cognitive, emotive and moral dimensions can be measured more appropriately through self-rating, while the behavioural domain is more relevantly measured by those in receipt of empathy. Thus, it would be unlikely for a single tool encompassing the same items to measure all aspects of empathy from the perspectives of self, raters and patients. Not surprisingly, inconsistent results and different ratings have been reported in the literature (La Monica 1981, 1987, Wheeler *et al.* 1996). The use of multiple measures of empathy to assess the many aspects of empathy is needed (Layton & Wykle 1990, Wheeler & Barrett 1994).

Recommendations for research and practice

A psychometrically and conceptually rigorous tool applied in the measurement of empathy in nursing research needs to be developed in the nursing context. Such a tool should be user-centred and cover all relevant domains reflecting the perspectives of users. The promotion of user involvement and provision of quality care appropriate to users' needs has been stressed in recent UK health policy documents (Department of Health 2005, Welsh Assembly Government 2005). A tool can give insight into users' views by involving them in the development of initial items, as well as during the validation process. Evidence should also be provided of a measure's potential to detect change when it does happen. A tool needs to be sensitive to change to evaluate training programmes aimed at developing empathy. Furthermore, feasibility has a high priority in routine practice. Evidence is needed that a measure can be used practically to assess empathy, either through self-assessment or by patient-rating in busy clinical settings. Such a tool needs to be easy for participants to complete and for researchers to administer.

CONCLUSION

There is no 'gold standard' tool to measure empathy in the nursing context, although the Empathy Construct Rating Scale scored the highest for quality assessment and is the most popular measure in nursing. Empathy measures need to cover all domains reflecting user perspectives and need to be tested with the relevant population in appropriate care settings. Advances in the empathy measurement in nursing research will assist the development of interventions to improve the quality of nursing care and training programmes aimed at promoting empathy. The development of user-centred and appropriately evaluated empathy measures is a critical step in achieving these.

REFERENCES

Alligood M.R. (2005) Rethinking empathy in nursing education: shifting to a developmental view. *Annual Review of Nursing Education* 3, 299-309.

Astrom S., Nilsson M., Norberg A., Sandman P. & Winblad B. (1991) Staff burnout in dementia care relations to empathy and attitudes. *International Journal of Nursing Studies* 28(1), 65-75.

Astrom S., Nilsson M., Norberg A. & Winblad B. (1990) Empathy, experience of burnout and attitudes towards demented patients among nursing staff in geriatric care. *Journal of Advanced Nursing* 15, 1236-1244.

Bailey S. (1996) Levels of empathy of critical care nurses. *Australian Critical Care* 9(4), 121-122, 124-127.

Barrett-Lennard G.T. (1962) Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs* 76(43), 1-36 (Whole No. 562).

Barrett-Lennard G.T. (1981) The empathy cycle: refinement of a nuclear concept. *Journal of Counseling Psychology* 28(2), 91-100.

Becker H. & Sands D. (1988) The relationship of empathy to clinical experience among male and female nursing students. *Journal of Nursing Education* 27(5), 198-203.

Beddoe A.E. & Murphy S.O. (2004) Does mindfulness decrease stress and foster empathy among nursing students? *Journal of Nursing Education* 43(7), 305-312.

Bowling A. (2002) *Research methods in health: Investigating health and health services*, 2nd edn. Open University press, Buckingham.

Carkhuff R.R. (1969a) *Helping and human relations: a primer for lay and professional helper.* Holt, Rinehart, and Winston Inc., Toronto.

Carkhuff R.R. (1969b) Helper communication as a function of helpee affect and content. *Journal of Counseling Psychology* 16(2), 126-131.

Carkhuff R.R. (1969c) The perception of the effects of teacher-counselor education: The development of communication and discrimination selection indexes. *Counselor Education and Supervision* 8, 265-272.

Centre for Reviews and Dissemination (2001) *Understanding systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews*, CRD report 4, 2nd edn. NHS CRD, York.

Daniels T.G., Denny A. & Andrews D. (1988) Using microcounseling to teach RN nursing students skills of therapeutic communication. *Journal of Nursing Education* 27(6), 246-252.

Davis M.H. (1980) A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology* 10, 85.

Department of Health (2005) *Creating a Patient-led NHS: Delivering the NHS Improvement Plan.* Department of Health, London.

Department of Health (2008) *Nursing quality to be measured for compassion of care.* http://nds.coi.gov.uk/Content/Detail.asp?ReleaseID=370990&NewsAreaID=2 (accessed 1 Aug 2008).

Evans G.W., Wilt D.L., Alligood M.R. & O'Neil M. (1998) Empathy: a study of two types. *Issues Mental Health Nursing* 19(5), 453-461.

Fields S.K., Hojat M., Gonnella J.S., Mangione S., Kane G. & Magee M. (2004) Comparisons of nurses and physicians on an operational measure of empathy. *Evaluation & the Health Professions* 27(1), 80-94.

Grange A., Bekker H., Noyes J. & Langley P. (2007) Adequacy of health-related quality of life measures in children under 5 years old: Systematic review. *Journal of Advanced Nursing* 59(3), 197-220.

Greenhalgh J., Long A.F., Brettle A.J. & Grant M.J. (1998) Reviewing and selecting outcome measures for use in routine practice. *Journal of Evaluation in Clinical Practice* 4(4), 339-350.

Gunther M., Evans G., Mefford L. & Coe T.R. (2007) The relationship between leadership styles and empathy among student nurses. *Nursing Outlook* 55(4), 196-201.

Hogan R. (1969) Development of an empathy scale. *Journal of Counseling and Clinical Psychology* 33(3), 307-316.

Hojat M., Fields S.K. & Gonnella J.S. (2003) Empathy: an NP/MD comparison. *Nurse Practitioner* 28 (4), 45-47.

Hojat M., Gonnella J.S., Nasca T.J., Mangione S., Vergare M. & Magee M. (2002) Physician empathy: Definition, components, measurement, and relationship to gender and specialty. *American Journal of Psychiatry* 159(9), 1563-1569.

Hojat M., Mangione S., Nasca T.J, Cohen M.J.M., Gonnella J.S., Erdmann J.B. & Veloski J.J. (2001) The Jefferson scale of physician empathy: development and preliminary psychometric data. *Educational and Psychological Measurement* 61(2), 349-365.

Husted J.A., Cook R.J., Farewell V.T. & Gladman D.D. (2000) Methods for assessing responsiveness: a critical review and recommendations. *Journal of Clinical Epidemiology* 53(5), 459-468.

Kliszcz J., Nowicka-Sauer K., Trzeciak B. & Nowak P. (2006) Empathy in health care providers--validation study of the Polish version of the Jefferson Scale of Empathy. *Advances in Medical Sciences* 51, 220-225.

Kuremyr D., Kihlgren M., Norberg A., Astrrm S. & Karlsson I. (1994) Emotional experiences, empathy and burnout among staff caring for demented patients at a collective living unit and a nursing home. *Journal of Advanced Nursing* 19, 670-679.

La Monica E.L. (1981) Construct validity of an empathy instrument. *Research in Nursing and Health* 4, 389-400.

La Monica E.L., Madea A.R. & Oberst M.T. (1987) Empathy and nursing care outcomes. *Scholarly Inquiry for Nursing Practice: An International Journal* 1(3), 197-213.

Lauder W., Reynolds W., Smith A. & Sharkey S. (2002) A comparison of therapeutic commitment, role support, role competency and empathy in three cohorts of nursing students. *Journal of Psychiatric Mental Health Nursing* 9(4), 483-491.

Layton J.M. (1979) The use of modelling to teach empathy to nursing students. *Research in Nursing and Health* 2, 163-176.

Layton J.M. & Wykle M.H. (1990) A validity study of four empathy instruments. *Research in Nursing & Health* 13, 319-325.

Linn R.L. & Slinde J.A. (1977) The determination of the significance of change between pre- and post-testing periods. *Review of Educational Research* 47(1), 121-150.

Marco G.A., Reynolds W.J., Fonfria V.C., Muñoz I.A., Bellver N.C., Fonfria V.R., Fusté R.D. & Zubicoa F.C. (2004) Process of cross-cultural adaptation of the Spanish version of the Reynolds W. empathy survey. *Revista de Enfermerfa* 27(12), 65-70.

Mehrabian A. & Epstein N. (1972) A measure of emotional empathy. *Journal of Personality* 40(4), 525-543.

Morse J.M., Anderson G., Bottorff J.L., Yonge O., O'Brien B., Solberg S.M. & McIlveen K.H. (1992) Exploring empathy: a conceptual fit for nursing practice? *Journal of Nursing Scholarship* 24(4), 273-280.

Murphy P.A., Forrester D.A., Price D.M. & Monagham J.F. (1992) Empathy of intensive care nurses and critical care family needs assessment. *Heart and Lung* 21(1), 25-30.

Olson J.K. (1995) Relationships between nurse-expressed empathy patient-perceived empathy and patient distress. *Journal of Nursing Scholarship* 27(4), 317-322.

Olson J. & Hanchett E. (1997) Nurse-expressed empathy, patient outcomes, and development of a middle-range theory. *Journal of Nursing Scholarship* 29(1), 71-76.

Palsson M.B., Norberg A., Hallberg I.R. & Bjorvell H. (1996) Burnout, empathy, and sense of coherence among Swedish district nurses before and after systematic clinical supervision. *Scandinavian. Journal of Caring Sciences* 10, 19-26.

Polit D.F. & Beck C.T. (2004) *Nursing research: Principles and methods,* 7th edn. Lippincott Williams & Wilkins, Philadelphia.

Reid-Ponte P. (1992) Distress in cancer patients and primary nurses' empathy skills. *Cancer Nursing* 15(4), 283-292.

Reynolds W.J. & Presly A.S. (1988) A study of empathy in student nurses. *Nurse Education Today* 8, 123-130.

Reynolds W.J, Scott B. & Jessiman W.C. (1999) Empathy has not been measured in clients' terms or effectively taught: a review of the literature. *Journal of Advanced Nursing* 30(5), 1177-1185.

Reynolds W. J. (2000) *The measurement and development of empathy in nursing.* Ashgate, Aldershot.

Roach K.E. (2006) Measurement of health outcomes: reliability, validity and responsiveness. *Journal of Prosthetics and Orthotics* 18(1s), 8-12.

Rogers C.R. (1957) The necessary and sufficient conditions of therapeutic personality change. *Journal of Counseling Psychology* 21(2), 95-103.

Russell I.T., Blasi Z.D., Lambert M.F. & Russell D. (1998) Systematic reviews and meta-analyses: opportunities and threats. In *Evidence-based fertility treatment* (Templeton A. & O'Brien P., eds), RCOG Press, London, pp.15-64.

Salovey P. & Mayer J.D. (1990) Emotional intelligence. *Imagination, Cognition and Personality* 9, 185-211.

Schutte N.S., Malouff J.M., Hall L.E., Haggerty D.J., Cooper J.T., Golden C.J. & Dornheim L. (1998) Development and validation of a measure of emotional intelligence. *Personality and Individual Differences* 25 (2), 167-177.

Streiner D.L. & Norman G.R. (2003) *Health measurement scales: A practical guide to their development and us*e, 3rd edn. Oxford University Press, Oxford.

Warner R.R. (1992) Nurses' empathy and patients' satisfaction with nursing care. *Journal of the New York State Nurses Association* 23(4), 8-11.

Watt-Watson J., Garfinkel P., Gallop R., Stevens B. & Streiner D. (2000) The impact of nurses' empathic responses on patients' pain management in acute care. *Nursing Research* 49(4), 191-200.

Welsh Assembly Government (2005) *Designed for life: creating world class health and social care for Wales in the 21st century.* http://www.wales.nhs.uk/documents/designed-for-life-e.pdf (accessed 18 Oct 2006).

Wheeler K. (1990) Perception of empathy inventory. In *Measurement of nursing outcomes: Measuring client self-care and coding skills of nursing outcomes, Vol. 4,* (Srickland O. & Waltz C., eds). Springer, New York, pp. 81-198.

Wheeler K. & Barrett E.A.M. (1994) Review and synthesis of selected nursing studies on teaching empathy and implications for nursing research and education. Nursing Outlook 42(5), 230-236.

Wheeler K., Marrett E.A.M. & Lahey E.M. (1996) A study of empathy as a nursing care outcome measure. *International Journal of Psychiatric Nursing Research* 3(1), 281-289.

Williams J.K., Skirton H., Reed D., Johnson M., Maas M. & Daack-Hirsch S. (2001) Genetic counselling outcomes validation by genetics nurses in the UK and US. *Journal of Nursing Scholarship* 33(4), 369-374.

Wilt D.L., Evans G.W., Muenchen R.A. & Guegold G. (1995) Teaching with Entertainment Films: An Empathic Focus. *Journal of Psychosocial Nursing and Mental Health Services* 33(6), 5-14.

Yu J. & Kirk M. (2008) Measurement of empathy in nursing research: systematic review. *Journal of Advanced Nursing* 64(5), 440-454.

Table 1: The seven criteria for quality appraisal of empathy measurement tools

Criteria	Description	Score		
		0	1	2
Validity	Construct validity and criterion validity	Low[1] One type of validity	Modest[2] One type of validity	High[3] Two types of validity
Reliability	Internal consistency, stability and equivalence	Low[1]	Modest[2] One type of reliability	High[3] Two or more types of reliability
Responsiveness	Tests for differences between individuals, factors associated with good outcome and treatment effect from group differences	None	Sensitive One type of test	Sensitive Two or more types of tests
Setting	Nursing, medical, counselling, or non-healthcare settings	Non-healthcare context	Medical or counselling context	Nursing context
User-centeredness	Whether and how to take users' views into account	Take no account of users' perceptions	Listen to users' views, but not ask them about how factors should be defined	Actively encourage users to specify what factors should be recorded and how they should be defined
Domain	Cognitive, emotional, moral and behavioural dimensions	One dimension	Two-three dimensions	All dimensions
Practicality & application	Whether it is easy to complete or widely used	Not easy to complete Used in a single nursing study	Need training for rating or administering Used in two-three nursing studied	Feasible and easy to compete Used in more than three nursing studies

[1] <0.5; [2] 0.5-0.75; [3] >0.75

Table 2: General information of the included empathy measures

Measure, reference	Origin, population	Domains	Items	Administration	Nursing studies using the measure
Barrett-Lennard Relationship Inventory - Empathy Understanding (Barrett-Lennard 1962)	USA Therapists (n=21) Clients (n=42)	Behavioural	16 items (+/-) 6-point scale +3=yes -3=no	Self-rating Client-rating	Olson 1995, Olson & Hanchett 1997
Carkhuff Indices of Discrimination & Communication (Carkhuff 1969a)	USA Counsellors Helpers in general	Cognitive Behavioural	16 scenarios 5-point scale 1=poor 5=good	Trained-rater-rating	Daniels *et al.* 1988
Emotional Empathy Tendency Scale (Mehrabian & Epstein 1972)	USA Undergraduate students (n=81-88)	Emotional	33 items (+/-) 7 subscales +4=very strongly agree -4=very strongly disagree	Self-rating	Gunther *et al.* 2007
Emotional Intelligence Scale (Schutte *et al.* 1998)	USA College students (n=23-346) General population	Emotional	33 items 5-point Likert scale 1=strongly disagree 5=strongly agree	Self-rating	Kliszcz 2006
Empathy Construct Rating Scale (La Monica 1981)	USA Nurses (n=173) Nursing graduate students (n=127) Patients (n=300) Professional peers (n=300)	Cognitive Behavioural	84 items (+/-) 6-point Likert scale +3=extremely like -3=extremely unlike	Self-rating Patient-rating Peer-rating	La Monica 1987, Daniels *et al.* 1988, Reynolds & Presly 1988, Astrom *et al.* 1990, 1991, Murphy *et al.* 1992, Warner 1992, Kuremyr *et al.* 1994, Bailey 1996, Palsson *et al.*

					1996
Hogan Empathy Scale (Hogan 1969)	USA General population	Cognitive Emotional Moral	39 true/false statements	Self-rating	Reynolds & Presly 1988, Evans *et al.* 1998, Gunther *et al.* 2007
Interpersonal Reactivity Index (Davis 1980)	USA Undergraduate students (n=427-1161)	Cognitive Emotional	28 items 5-point Likert scale 0=does not describe me well 4=describe me well Four subscales	Self-rating	Becker & Sands 1988, Beddoe & Murphy 2004, Kliszcz 2006
Jefferson Scale of Physician Empathy (Hojat *et al.* 2001)	USA Physicians (n=55) Residents (n=41) Medical students (n=193)	Cognitive	20 items 7-point Likert Scale 1=strongly disagree 7=strongly agree	Self-rating	Hojat *et al.* 2003, Fields *et al.* 2004, Kliszcz 2006
Layton Empathy Test (Layton 1979)	USA Nursing students (n=56)	Cognitive Behavioural	Two forms (I & II) Part 1: 12 true/false items Part 2 and 3: 6 2-choice items, selecting the most/least empathic response	Self-rating	Wilt *et al.* 1995, Wheeler *et al.* 1996, Evans *et al.* 1998
Perception of Empathy Inventory (Wheeler 1990)	USA Patients (n=81)	Behavioural	33 items True/false statements 4-point Likert scale 1=not at all true 4=very true	Patient-rating	Wheeler *et al.* 1996
Reynolds Empathy Scale	UK Nurses (n=32-103)	Behavioural	12 items 7-point Likert scale	Trained-rater-rating	Lauder *et al.* 2002

(Reynolds 2000)			0=never like 6=always like		
Visual Analogue Scale (Wheeler *et al.* 1996)	USA Nursing students (n=82)	Behavioural	3 bipolar statements 100mm scale	Clinical-teacher-rating	Wheeler *et al.* 1996

USA = United States of America
UK = United Kingdom

Table 3: Psychometric properties of the included empathy measures

Measure	Item generation, content/face validity	Construct, criterion validity	Reliability	Responsiveness	Comments
Barrett-Lennard Relationship Inventory - Empathy Understanding	Based on Rogers' (1957) theory and Bown's (1954) Relationship Sort **Content validity** 5 judges (counsellors)	**Construct validity** Comparison between expert and non-expert therapists (significantly different, no P value reported) Agreement between expert therapist-client pairs and non-expert therapist-client pairs ($P<0.01$) **Predictive validity** Correlation between therapist empathy and client therapy outcome improvement ($P<0.05$)	**Subscale inter-correlations** Mean r=0.45, clients Mean r=0.65, therapists **Split-half reliability** r=0.86, clients r=0.96, therapists **Test-retest reliability** r=0.89, over 4 weeks	Not reported	Easy to administrate; a subscale of the Relationship Inventory; assessing therapist-client relationships, but not nurse-patient relationships; measuring client-perceived empathy; small sample size; low reliability
Carkhuff Indices of Discrimination & Communication	Not stated	**Construct validity** Compared seven subgroups (significantly different, no P value reported)	**Test-retest reliability** r=0.93, r=0.95 **Inter-rater reliability** r=0.89	Evaluating a training programme	Assesses cognitive appreciation of empathy, rather than empathy ability per se; counsellor-client or helping relationships; most scenarios were female-focused; good reliability; some responsiveness
Emotional Empathy	Selected from a larger pool of	**Construct validity** Correlation between empathy	**Subscale inter-correlations**	Not reported	Tested for undergraduate students; not relevant in

Tendency Scale	items (not specified)	and aggression (β= -0.21, P=0.05), helping behaviour (β=0.31, P=0.05) and gender (r= -0.42) **Discriminant validity** Compared with the Social Desirability Scale (r=0.06, non-significant)	Rs>0.30, P=0.01 **Split-half reliability** r=0.84		healthcare settings; measuring cognitive appreciation of empathy; low reliability
Emotional Intelligence Scale	Based on the theoretical model of emotional intelligence	**Construct validity** Factor analysis (one factor of 33 items) **Convergent validity** Compared with 6 conceptually relevant measures (r= -0.37 to 0.68, P<0.02 at least) Between-group differences [t(37)=2.35, p<0.012; t(25)=1.86, P<0.035] Gender comparison [t(327)=3.29, P<0.001] **Discriminant validity** Compared with Scholastic Assessment Test (r= -0.06, non-significant) and NEO Personality Inventory [r(22)=0.54, P<0.009, only 1 of the 6 dimensions] **Predictive validity** r=0.32, P<0.01	**Cronbach's alpha** α=0.87, college students α=0.90, respondents from various settings **Test-retest reliability** r=0.78, over two weeks	Not reported	Not developed in health care settings; small sample size; low validity; moderate reliability
Empathy	Generated by	**Construct validity**	**Cronbach's alpha**	Not reported	From females'

Construct Rating Scale	female graduate students from psychology and nursing **Face/content validity** A panel of 3 experts and students	Factor analysis (r=0.92, between well-developed and lack-of-empathy items) **Convergent validity:** not evident **Discriminant validity** r=0.20, P<0.001 (empathy-self and empathy-client) r=0.10, P<0.05 (empathy-self and empathy-peer) r=0.06, P>0.05 (empathy-peer and empathy-client)	α=0.97 **Split-half reliability** Form A (well-developed empathy items): r=0.89 Form B (lack-of-empathy items): r=0.96 **Test-retest reliability** Form B: α=0.98		perspectives; too lengthy; not addressing nurse-patient interactions and nurses' actual experiences; low validity; good reliability
Hogan Empathy Scale	An item analysis of the responses of high-rated and low-rated empathy groups **Content validity** Psychology students, psychologists and lay population	**Construct validity** Group and gender comparison in 11 male groups and 3 female groups; high school students (P=0.05, P<0.001) **Concurrent validity** Compared with Q-sort-derived empathy ratings (r=0.62, general population; r=0.39, medical school applicants) and 'social acuity' scale (Mean r=0.58, general population; Mean r=0.42, medical school applicants)	**Reliability coefficient** (Kuder-Richardson 21) r=0.71 **Test-retest reliability** r=0.84, over two months	Not reported	Not measuring expressed empathy; tested in psychology students, psychologists and lay population; not specific for health settings; comprehensive tested with moderate to high reliability and validity
Interpersonal Reactivity Index	Generated with some borrowed or adapted from	**Construct validity** Factor analysis (4 factors) Gender comparison (P<0.01)	**Cronbach's alpha** a=0.70-0.78 (females) a=0.75-0.78 (males)	Not reported	Not for nurses; not specific to professional helping or clinical situations; easy to

	other measures **Content validity** Students from introductory psychology classes		**Subscale inter-correlations** Low or non-significant **Test-retest reliability** r=0.61-0.79, males r=0.62-0.81, females, over 60 to 75 days		administer; moderate reliability and validity
Jefferson Scale of Physician Empathy	Based on a literature review **Face validity** 100 physician, Delphi method	**Construct validity** Factor analysis (4 factors) Gender comparison (t=2.41, P<0.05) **Convergent validity** Compared with related measures (r=0.12-0.56, P<0.01) **Discriminant validity** Compared with unrelated scales (r=0.05-0.11, statistically non-significant)	**Cronbach's alpha** α=0.87, residents α=0.89, medical students	Not reported	In medical settings; no behavioural component, no patients' views; low validity; good reliability
Layton Empathy Test	Based on a literature and professional experience **Content validity** Nursing faculty members	**Construct validity** Compared with the Carkhuff Empathic Understanding Scale (r=0.46, P<0.01) and the Barrett-Lennard Empathy Relationship Inventory (statistically non-significant, no P value reported)	**Reliability coefficient** (Kuder-Richardson 20) r=0.24, Form I; r=0.26, Form II	Evaluating the use of modelling to teach empathy in treatment and control groups	For nursing students; easy to administer; tested in small number of respondents in each of the 5 subgroup (from 3 to 8); low reliability and validity; report on responsiveness

Perception of Empathy Inventory	Based on items from BLRI empathy subscale and a review of nursing literature **Face/content validity** 2 professors in psychiatric nursing and 4 patients	**Construct validity** Correlation between nurse empathy and patient anxiety (r=0.52, P=0.008) Correlation with demographic variables (Non-significant)	**Cronbach's alpha** α=0.94 **Item-total correlations** r=0.21-0.83	Not reported	For patients; validated in small number of patients; moderate validity; low to high reliability
Reynolds Empathy Scale	Developed according to clients' perceptions of effective and ineffective interpersonal behaviours **Face/content validity** A panel of experts	**Concurrent validity** Compared with Empathy Construct Rating Scale (r=0.85, P<0.001)	**Cronbach's alpha** α=0.90, nurses **Internal discrimination** Phi coefficient, most values ranged from around 0.80 and above, p<0.001 **Test-retest reliability** r = 0.90, p<0.001, over 2-4 weeks **Inter-rater reliability** Initial: 25-33% Final: 41.6-91.6%	Evaluating the effect of a training programme	The only empathy tool developed in the United Kingdom and one of the few ones developed for nurses; patients not involved in its assessment; good validity and reliability, report on responsiveness
Visual Analogue Scale	Not stated	**Concurrent validity** Compared with Layton Empathy Test (r=0.26, P=0.05)	**Cronbach's alpha** α=0.68, nursing students	Not reported	For nursing students; low validity; moderate reliability

Table 4: Measures against the seven criteria for quality appraisal of each of the empathy measurement tools

	BLRI	CIDC	ECRS	EIS	EETS	HES	IRI	JSPE	LET	PEI	RES	VAS
Validity	2	1	1	1	1	1	1	1	-	1	1	-
Reliability	1	2	2	2	1	1	1	1	-	1	1	1
Responsiveness	-	2	-	-	-	-	-	-	1	-	1	-
Setting	1	1	2	-	-	-	-	1	2	2	2	2
User-centeredness	-	-	1	-	-	-	-	-	-	-	1	-
Domain	-	1	1	-	-	1	1	-	1	-	-	-
Practicality & application	-	-	1	-	-	1	1	1	1	-	-	-
Quality score (out of 14)	4	7	8	3	2	4	4	4	5	4	6	3

BLRI: Barrett-Lennard Relationship Inventory - Empathy Understanding
CIDC: Carkhuff Indices of Discrimination & Communication
ECRS: Empathy Construct Rating Scale
EETS: Emotional Empathy Tendency Scale
EIS: Emotional Intelligence Scale
IRI: Interpersonal Reactivity Index
HES: Hogan Empathy Scale
JSPE: Jefferson Scale of Physician Empathy
LET: Layton Empathy Test
PEI: Perception of Empathy Inventory
RES: Reynolds Empathy Scale
VAS: Visual Analogue Scale

Figure 1: Flow diagram of the search process

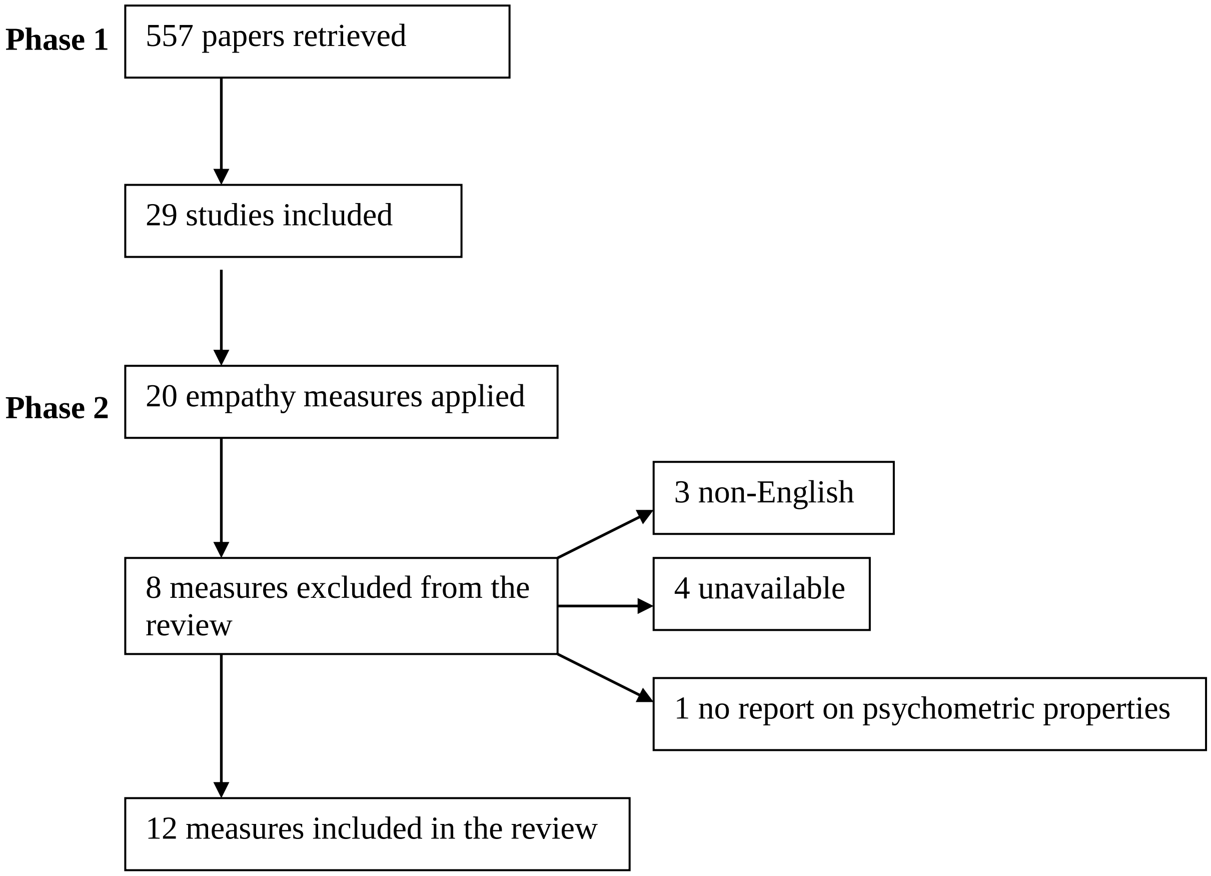
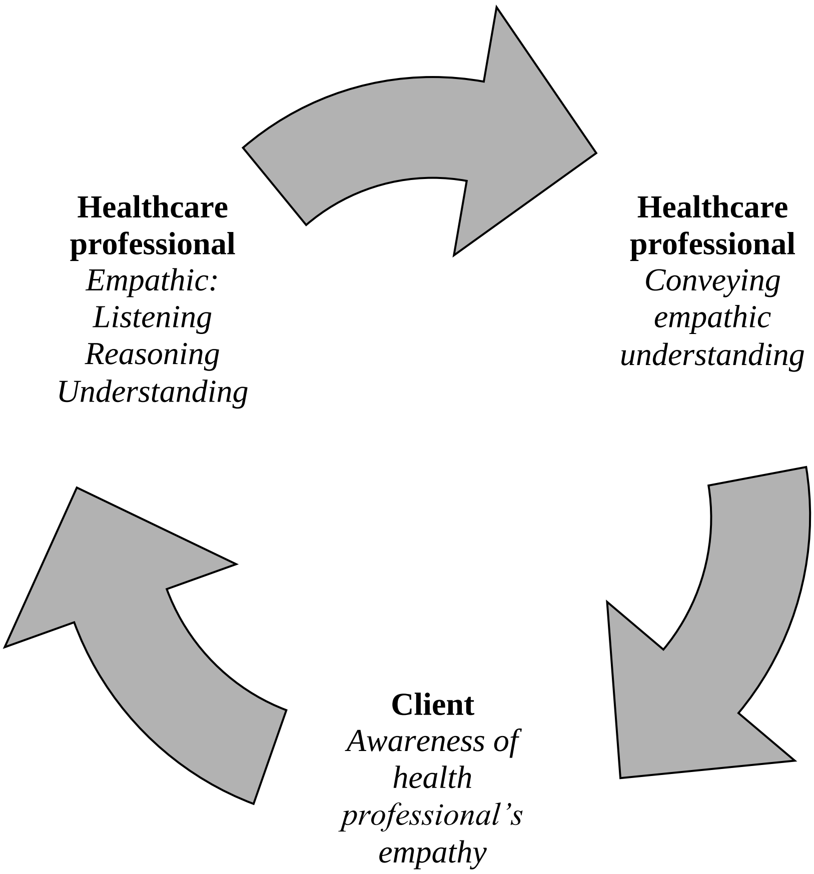

Phase 1
557 papers retrieved
29 studies included
Phase 2
20 empathy measures applied
3 non-English
8 measures excluded from the review
4 unavailable
1 no report on psychometric properties
12 measures included in the review

Figure 2: The empathy cycle of Barrett-Lennard (1981)


Healthcare professional
Empathic:
Listening
Reasoning
Understanding
Healthcare professional
Conveying empathic understanding
Client
Awareness of health professional's empathy