Rigour, reliability and validity in qualitative research

T. Long and M. Johnson

This article addresses issues relating to rigour within qualitative research, beginning with the need for rigour at all in such studies. The concept of reliability is then analysed, establishing the traditional understanding of the term, and evaluating alternative terms. A similar exploration of validity and proposed alternatives follows. It is suggested that there is nothing to be gained from the use of alternative terms which, on analysis, often prove to be identical to the traditional terms of reliability and validity. Alternative or novel means of addressing these concepts in interpretive research are, however, welcomed. A review of some of the strategies available for the pursuit of reliability and validity in qualitative research is undertaken. These are clearly identified as means to establish existing criteria and are found to have variable value.

Keywords: rigour, reliability, validity, qualitative research

THE NEED FOR RIGOUR IN RESEARCH

There is general agreement that all research studies must be open to critique and evaluation. Failure to assess the worth of a study – the soundness of its method, the accuracy of its findings, and the integrity of assumptions made or conclusions reached – could have dire consequences. Ambiguous or meaningless findings may result in wasted time and effort, while findings which are simply wrong could result in the adoption of dangerous or harmful practices. Evaluation of studies, then, is an essential pre-requisite of the application of findings. Traditionally, such evaluation has centred on assessment of reliability and validity. However, while these terms have distinct meanings which relate well to other concepts and assumptions within the logical-positivist paradigm, their use in qualitative work has been questioned. A variety of positions is to be found. These include dismissing any attempt to establish rigour; using existing terms and criteria in the traditional manner; using existing terms and criteria with modification to their interpretation; and rejecting traditional terms and criteria while substituting new terms and criteria. It the last of these which is challenged here.

RELIABILITY

Taking two commonly used sources for the traditional understanding of reliability, this concept can be described as 'the consistency or constancy of a measuring instrument' (LoBiondo-Wood & Haber 1998, p. 558), or 'the degree of consistency or dependability with which an instrument measures the attribute it is designed to measure' (Polit & Hungler 1995, p. 651). Hammersley (1992, p. 67) suggests that reliability 'refers to the degree of consistency with which instances are assigned to the same category by different observers or by the same observer on different occasions.' Surprisingly, in view of their different philosophical approaches, there is little disagreement among these definitions. The first two indicate the detached nature of the researcher, while the third acknowledges the active involvement of the researcher, but all relate to confidence in data collection.

The traditional understanding of reliability focuses on standardizing data collection instruments (Mason 1996, p. 24). However, 'this is premised on the assumption that methods of data generation can be conceptualised as tools, and can be standardised, neutral and non-biased' (Mason 1996, p. 145). While this may be acceptable for quantitative methods, though Hammersley (1992) questions even this, the non-standardization of qualitative methods and the determination to seek greater validity through retention of context makes it impossible in qualitative work. Brink (1991, p. 176) proposes three tests of reliability for qualitative work, each to be used as is appropriate for specific studies. *Stability* is established when asking identical questions of an informant at different times produces consistent answers. *Consistency* refers to the

Tony Long PhD, RN
Senior Nursing Lecturer, School of Healthcare Studies, Baines Wing. University of Leeds PO Box 214 Leeds LS2 9UT, UK

Martin Johnson PhD, RN
Professor of Nursing, Dept of Acute & Critical Care Nursing, Greenbank Building, University of Central Lancashire, Preston PR1 2HE, UK

Correspondence to: TL at University of Leeds. Tel: +44(0) 113 233 1297; Fax: +44 (0) 113 233 1204; E-mail: a.j.long@leeds.ac.uk

integrity of issues within a single interview or questionnaire, so that a respondent's answers on a given topic remain concordant. *Equivalence* is tested by the use of alternative forms of a question with the same meaning during a single interview, or by concurrent observation by two researchers. These approaches appear to do no more than apply standard approaches of replicability and inter-rater reliability to qualitative interviews or observation studies, seeking inappropriately to standardise highly variable data collection methods.

There seems to be a growing popular movement within qualitative circles to insist that 'dependability' is a more appropriate term than reliability for qualitative research (Sandelowski 1986; Hall & Stevens 1991; Robson 1993; Koch 1994). The origin of this movement was the work of Guba and Lincoln (1985), but the authors' thoughts were refined in their contribution of 1989. In this they explain that 'dependability is parallel to the conventional criterion of reliability, in that it is concerned with the stability of data over time' (Guba & Lincoln 1989, p. 242). However, the concern at the root of dependability is the same as that for reliability: to ensure that data collection is undertaken in a consistent manner free from undue variation which unknowingly exerts an effect on the nature of the data. A specific strategy of dependability audit is suggested by which to enhance or demonstrate dependability, but this is clearly declared by Guba and Lincoln (1989, p. 242) as a means to an end rather than a criterion in itself.

Some accounts lack any suggestion of explanation of the need for an alternative criterion. Robson (1993, p. 405), for example, provides no description or explanation of dependability other than to state that 'dependability is analogous to reliability.' The account continues to describe how dependability is established through auditing of the decision trail. As Hammersley (1992) suggests is often the case, the focus is upon means rather than criteria. A frequently quoted article by Koch (1994), which in fairness is intended to be a discussion only of the decision trail, makes no attempt to define or describe any of Guba and Lincoln's alternative criteria, but is often cited as evidence of the need for such terms (e.g. Holloway & Wheeler 1996).

The inescapable conclusion is that the terms considered above have the same essential meaning, and nothing is to be gained from clouding the issue with alternative labels for what have been argued to be identical concepts. Reliability, recognized as pertaining to the stability of data-collection measures, remains an important notion. Rather than attempting to hide behind a smokescreen of synonyms, perhaps interpretive researchers ought simply to accept that reliability is unlikely to be a demonstrable strength of their work. Although efforts may be made to enhance a study's reliability, in most cases the nature of the data and the sample make this practically hopeless.

VALIDITY

In quantitative terms, validity is taken to mean 'the determination of whether a measurement instrument actually measures what it is purported to measure' (LoBiondo-Wood & Haber 1998, p. 561), or 'the degree to which an instrument measures what it is intended to measure' (Polit & Hungler 1995, p. 656). Hammersley (1992, p. 69) provides a qualitative perspective: 'An account is valid or true if it represents accurately those features of the phenomena that it is intended to describe, explain or theorise.' As with definitions of reliability, there seems little discrepancy between these perspectives.

While in quantitative studies validity is concerned specifically with avoidance of type I and type II errors, Silverman (1993, pp 144, 155) argues that the refutability proposed by Popper (1979) is the standard approach to testing the validity of any research findings. Hammersley (1992, p. 69) accepts that no knowledge can be counted as certain, and the best that we can do is to seek means of judging claims to knowledge in terms of their likely truth. These means are laid out as considering the plausibility of the claim, the credibility of the claim, and the weight of evidence for each of these. Silverman, noting that claims to credibility may be no more than a reflection of uncritical public ignorance, agrees that the only useful option is to consider the quality of the evidence which upholds the claims. Hammersley (1992, pp 70–72) refers to the realities of life and differing degrees of need for confidence dependent upon the significance of the claim. He suggests that some claims are within our common experience, leave little room for error on the part of the researcher, or are in no way central to the main issues of the study. These may be accepted without much concern. Other claims allow more room for misinterpretation by the researcher and therefore require stronger evidence. Where a claim is of particular significance to the study the most convincing evidence is called for. Although Silverman (1993, pp 155–156) shuns all but the most demanding of these, it is clear that Hammersley is committed to any claims (including those based on plausibility or credibility) to knowledge being supported by evidence. It is the nature of the claim and its centrality to the study which decide the degree of evidence required.

Validity is normally established through consideration of three main aspects: content validity, criterion-related validity and construct validity. The first of these depends largely on sampling and careful construction of the instrument and refers to the degree to which the entirety of the phenomenon under investigation is addressed. A sub-set of this is the weak concept of face validity which assures only that the instrument and findings appear to be thorough and accurate to reputedly knowledgeable reviewers. Criterion-related validity is concerned with comparison of the instrument and findings

with an established standard to determine the correlation between measured performance and actual performance. Finally, construct validity is associated with consideration of the proximity of the instrument to the construct in question.

As with reliability, those who accept the need for tests of validity in qualitative research commonly insist on the use of alternative terms, often asserting that these explain alternative concepts. Guba and Lincoln (1989, pp 236–237) adopt the term 'credibility' instead of validity. They argue that validity refers to the naïve reality of positivism and an attempt to establish 'isomorphism between findings and objective reality.' Their substitution involves replacing this with 'isomorphism between constructed realities of respondents and the reconstructions attributed to them' (Guba & Lincoln 1989, pp 236–237). The only difference between the terms is the presumed objective reality of positivism and the constructed realities of constructivism. The underlying concept appears to be identical: to match what is reported by the researcher to the phenomenon under investigation. As suggested by Hammersley (1992, p. 67), the notable difference lies more in the means to establish achievement of the criterion rather than in the criterion itself. However, the assertion of the need for a new criterion is accepted in a fairly unquestioning manner in a number of texts directed specifically at nurses (e.g. Holloway & Wheeler 1996).

Further claims are often made for feminist research. Hall and Stevens (1991), for example, explain ten measures of 'adequacy'. Adequacy is preferred as a term to validity because it is held that reliability and validity cannot be separated in feminist research as they are in other approaches. In fact, they also explain and employ the term 'dependability', which, it is argued above, is identical to reliability, and they discuss this separately. The origin of the assumption that reliability and validity are unconnected in traditional studies is not made clear, although Hammersley (1992, p. 66) provides more argument for this. He notes that 'the findings of a study are either valid or they are not … The distinction between internal and external validity is fundamentally misleading.' Adequacy is explained as follows. 'Results are adequate if analytic interpretations fairly and accurately reflect the phenomena that investigators claim to represent' (Hall & Stevens 1991). This appears to concur with the definitions of validity already considered, thus calling into question the assertion that 'adequacy' is in some way different to, or more than, validity. In seeking means to check credibility Hall and Stevens (1991) demand a depth of trust, intimacy of setting, and sensitivity to the language and life-style of the respondent in addition to sufficient length and frequency of contact in order to demonstrate 'rapport'. The purpose of this is clearly to ensure that the full story unfolds and thereby to ensure what is otherwise known as content validity.

'Coherence', say Hall and Stevens (1991), 'can be recognised in the consistency of the whole with its constituent parts.' Thus, the plausibility of the findings is reliant upon the interpreted, analysed data being recognisably drawn from the raw data, and is demonstrated partly by evidence of stringent efforts to ensure this. In other work (Silverman 1993), this requirement is held to be of considerable importance and is addressed, for example, by the process of analytic induction. Although not expressed by Hall and Stevens, there must also be acknowledgement of the need for the data to have been collected in a valid manner. These concerns to ensure that the findings appear believable and that they are representative of the mass of raw data (and, therefore, in turn, of the original phenomenon) are directly comparable respectively to face and content validity.

Two other criteria proposed by Hall and Stevens (1991) are more problematic: 'complexity' and 'consensus'. They justifiably demand that findings are context-specific and remain sufficiently complex to reflect accurately the true nature of complex phenomena. This obviously expresses a commitment to content validity. However, they also note that 'congruence among behavioural, verbal and affective elements of particular observations, verbal responses, and written records helps to support the presence of consensus.' The suggestion is that credibility is enhanced when triangulated data concur, and that areas where congruence occurs should be given priority. This must automatically limit the breadth of data which is deemed to be acceptable, thereby stripping the complexity of the data through discarding contextual material. Silverman warns of this danger when utilizing triangulation. All accounts or observations could be true within their own context, and 'counterposing different contexts ignores the context-bound and skilful character of social interaction' (Silverman 1993, pp 156–158). Hall and Stevens (1991) admit that 'inconsistency among participant accounts does not invalidate their perceptions, but instead illustrates the variety of women's thoughts, actions and feelings.' They seem to recognize the apparent disparity between complexity and consensus, but offer no solution.

'Relationality' is the term used by Hall and Stevens to address the group co-operation with other researchers and participants which enables larger samples, longer studies and more critical reflection. Wider sampling, larger numbers and greater depth of critical analysis of alternative explanations are principles held dear by positivists, too, and are generally sought in order to enhance content validity.

'Reflexivity' is the term used by Hall and Stevens (1991) for the recognition of the need to incorporate the subjective value of the researcher's

feelings and attitudes into consideration of the findings. A word of caution is offered by Hammersley (1992, p. 142), warning that 'there is a danger of various substantive values being smuggled in under the disguise of the formal value of reflection.' Despite this, Koch (1994) notes that giving sufficient detail of the researcher's and participants' involvement, together with existing knowledge and beliefs, allows estimation of adherence to, or departure from the theoretical construct under investigation. Such concern is usually addressed as construct validity. Although claimed as a criterion for the new concept of adequacy, reflexivity is really presented as a means to the accomplishment of the existing criterion of construct validity.

Hall and Stevens (1991) view credibility, however, from a different perspective. For them a study 'is credible when it presents such faithful interpretations of participants' experiences that they are able to recognise them as their own.' They appear to seek credibility in the eyes only of the participants rather than external reviewers, though this is, of course, compatible with the principle proposed above by Hall and Stevens that what women say is inherently valid. For much qualitative research, there is little possibility of finding a standard against which to test the study findings and respondent validation is utilized instead. In this, the respondents themselves provide such a standard. What is apparently proposed here is precisely that, and thereby criterion-related validity is being sought.

Two more criteria of 'honesty and mutuality' are held to relate specifically to feminist research. It is assumed that women are inherently honest and able to ignore the effects of being studied, and it is then sufficient to treat women subjects as equals for 'the subjective validity of participants' statements, affects and behaviours' to be preserved (Hall & Stevens 1991). This appears to be an unsupported assertion that women guard a self-evidently valid construct of a given phenomenon, but in any case admits concern for construct validity. This is reinforced by 'naming': since 'a study on feminist principles is adequate if the active voices of women participants are heard in the research account' (Hall & Stevens 1991). Although this is clearly no guarantee that the researcher heard, recorded or interpreted those voices accurately, it once again confirms the value of construct validity.

Hall and Stevens' (1991) final criterion of 'relevance', which asserts that research is not valid if it fails to further the struggle for women's emancipation, is clearly not associated with the concepts of validity outlined above, but may be a specific example of what Hammersley (1992, pp 72–77) refers to as 'relevance'. While the assumptions made and terminology employed by Hall and Stevens are merely examples of alternatives to traditional terms of validity, it can be seen that there is no substantive distinction in meaning or concern.

THE MEANS TO ESTABLISH RIGOUR

Hammersley (1992, p. 67) suggests that there is great confusion between criteria of rigorous research and the means by which the criteria may be evaluated in qualitative research. Once disentangled the more common of these include, for reliability, audit of the decision trail and triangulation. For validity, the means include self-description and reflective journal-keeping; respondent validation; prolonged involvement; persistent observation; peer debriefing; and triangulation.

Self-description and reflective journal

Reflection is an essential part of qualitative research. Porter (1993) sees reflexivity as researchers reflecting on their own beliefs in the same manner as they examine those of their respondents. These beliefs and values are made explicit and taken into account so that 'rather than engaging in futile attempts to eliminate the effects of the researcher, reflexive researchers try to understand them' (Hammersley & Atkinson 1995, p. 18).

Respondent validation (member check)

Brink (1991) suggests the use of respondent validation (Bloor 1978) to ensure stability. Checking the results on completion of data collection or of the whole study with the respondents would, it is alleged, meet the requirements of diachronic reliability (stability over time). However, significant elements of raw data are made up of field notes, observation of non-verbal signs, and recognition of unconscious changes in tone and emphasis. These may not be acknowledged or accepted later by the respondent. Furthermore, even if validation is to be undertaken at the end of the study the time lapse involved is unlikely to be sufficient to demonstrate stability in a meaningful way. Alternatively, supposing that respondent validation were to be undertaken two or three years after a study, it might be expected to provide more security regarding the stability of the findings. There would, however, be considerable problems associated with such an attempt, such as respondent morbidity, lack of access, and alteration of the respondent's situation and views, perhaps even as a result of participation in the study. 'Member check' has been used an alternative term for respondent validation. However, while Schein (1987, p. 51) suggests that this relates to researchers being able to pass themselves off as a member of the studied group (accepted by Bloor (1997) as one of three possibilities), the most common understanding of this would appear to be identical to the principles of respondent validation as above. The accuracy of

the findings is checked with members of the studied group.

Both Hammersley and Atkinson (1995, p. 227) and Mason (1996, pp 151–152) warn against placing too much faith in the results of respondent validation. The former note that 'we cannot assume that anyone is a privileged commentator on his or her own actions, in the sense that the truth of their account is guaranteed' (Hammersley & Atkinson 1995, p. 229). Participants' memory may fail, they may be unconscious of some of the non-verbal clues that they transmit which forms part of the data, or they may simply (consciously or unconsciously) deny less attractive aspects of their behaviour. Mason (1996, p. 147) adds that each individual respondent has no true insight into the experiences of other participants. Examples of all of these problems are provided by Bloor (1997, pp 41–48), who is emphatic that 'member validation is a many-splendoured thing, but it is not validation.' Hammersley and Atkinson (1995, p. 230) conclude that 'Such feedback can be highly problematic. Whether respondents are enthusiastic, indifferent, or hostile, their reactions cannot be taken as direct validation or refutation of the observer's inferences.'

A further issue relating to this is that respondent validation is normally conducted by the researcher. It might add strength to the process if the check were to be undertaken by a third party. The difficulties associated with this would be the lack of rapport with the respondents (likely to affect their responses), and the possibility of re-interpreting the findings during the process. An intermediate position might be for the researcher to present the findings to the respondents, either by telephone follow-up to written details or by means of a simple questionnaire, and then for a third party to undertake analysis and comparison with the study findings. While this might go some way to answering accusations of researchers sitting in judgement on themselves, it would do nothing to address the main problems highlighted by Hammersley and Atkinson or by Mason. While respondent validation may be a useful addition to the means of assessing the rigour of a study, the results, whether supportive or not, must be treated with caution.

Prolonged involvement and persistent observation

Kirk and Miller (1986, pp 30–31) argue that prolonged involvement in a community under research enhances sensitivity 'to discrepancies between the meanings presumed by the investigator and those understood by the target population.' Simply being in the respondent's environment enhances the likelihood of their meaning emerging and being recognised. It is a means to enhance validity, then, if the researcher can spend a significant length of time in contact with respondents individually and with the topic generally. This allows time for emerging concepts to develop and for potential implications to be recognized. It also allows for more opportunities to test out tentative explanations. Guba and Lincoln (1989, p. 237) recommend prolonged engagement in order to build trust and overcome the difficulties presented by perverse constructions and misinformation on the part of respondents. Persistent observation enhances the effect of this involvement by enabling the researcher 'to identify those characteristics and elements in the situation that are most relevant to the problem or issue being pursued and to focus on them in more detail' (Guba & Lincoln 1989, p. 237).

Peer debriefing

Robson (1993, p. 404) briefly describes peer debriefing as 'exploring one's analysis and conclusions to a colleague or other peer on a continuous basis.' He suggests that being explicit in formulating something for presentation to a peer fosters subsequent credibility. However, this is the limit of the explanation. Similarly, Holloway and Wheeler (1996, p. 165) mention only that supervisors have a key role with research students to ensure rigour in their studies. Peer debriefing may be pursued in numerous forms. One of these is to discuss the emerging findings at intervals with knowledgeable colleagues, a second to present and defend method and findings at national research conferences, and a third to present the findings and implications to interested groups. Review with colleagues is intended to stimulate consideration and exploration of additional perspectives and explanations at various stages of the process of data collection and analysis. In particular, it is aimed at preventing premature closure of the search for meaning and patterns in the data. Presentation at research conferences is a recognized means of submitting method and findings to other researchers so as to attract and answer to critical comment. This is the process which Hammersley (1992) refers to in his discussion of levels of evidence to substantiate claims to knowledge. Presenting findings and implications to interested users offers similar opportunities but with particular emphasis on the relevance of the study.

Triangulation

Triangulation may take several forms, but commonly refers to the employment of multiple data sources, data collection methods, or investigators. In general, the purpose of this would be to reduce the disadvantages inherent in the use of any single source, method or investigator. Kirk and Miller (1986, p. 30) note the usual concerns about avoiding type 1 and type 2 errors, but introduce concern for

'type 3' errors since 'asking the wrong question actually is the source of most validity errors.' Triangulation of method is held to be an effective device to prevent this. However, Hammersley and Atkinson (1995, pp 231–232) warn against the assumption that triangulation provides evidence that some data are true while other data is false. Observations from differing sources or resulting from different methods may well be observations of a different phenomenon. This alteration may not be immediately apparent. Although the new question may not be 'the wrong question', it could well be a different one. 'If it is accepted that there are horses for courses,' suggests Bloor (1997, p. 38), 'and that, for any given topic, there will be one best method of investigation, then triangulation may be said to involve juxtaposing findings gathered by the best available method with findings generated by an inferior method.' In this, he holds that triangulation may illuminate different perspectives on the problem but does not provide any test of validity. If there is a place for triangulation, Hammersley and Atkinson (1995, p. 232) emphasize that it should be 'a matter not of checking whether data are valid, but of discovering which inferences from those data are valid.'

Audit of the decision trail

This technique, first proposed by Sandelowski (1986), involves the presentation of details of all sources of data, collection techniques and experiences, assumptions made, decisions taken, meanings interpreted, and influences on the researcher. This was not an original idea, however. The notion of honestly declaring the origins of value-laden concepts and publicly acknowledging potentially researcher-centred perspectives was supported by Myrdal (1970, p. 43). The purpose of declaring the decision trail is to allow others to decide on the worth of the study by following the trial taken and comparing it with their own conclusions made from the same information. It is a demonstration of the degree to which the researcher has remained true to the data and to the boundaries of the sample. There is some obvious correlation between this and replicability, the possibility of duplicating a study to determine if earlier results are repeated. The most practical difference would appear to be that the former involves theoretical activity while the latter requires active repetition of the study in question. Each tests whether assumptions are justifiable and data collection has been undertaken rigorously and reported accurately. However, Sandelowski's proposal allows for the uniqueness of each situation and recognises the nature of the claim as one of a perspective on rather than reproduction of the phenomenon. Kirk and Miller (1986, pp 55–56) include the recording of field notes in this sort of strategy. They note the importance of retaining socially undesirable or irresponsible entries and distinguishing verbatim respondent items from researcher interpretations. The decision trail may be displayed in the discussion of the method used (and particularly issues relating to the sample), in the detailed exposition of data analysis, and within the discussion of findings.

CONCLUSION

There is a clear imperative for rigour to be pursued in qualitative research so that findings may carry conviction and strength. Reliability is a concept applicable to qualitative studies without the need for alternative terminology. While the catalogue of existing strategies for assessment of reliability may not normally be appropriate in interpretive studies, reliability in interpretive work can be assessed and presented using the existing terms but employing different means. In most cases it will be exceptionally difficult to demonstrate reliability. However, this is to be expected, as perfect validity is the sole guarantor of reliability. Validity has been found to be the quintessential element of qualitative research, sharing the same meaning and terminology as traditional approaches, but bearing especially great significance. However, alternative strategies for assessment and assurance of validity may be required for such studies. Where existing terms and concepts can be used without prejudice to data and findings from qualitative studies alternatives are superfluous and serve only to alienate those with differing priorities or stances. For clarity and utility identical terms and concepts should be addressed. The need is not for new criteria or novel terms but for different means of addressing existing criteria.

REFERENCES

Bloor M 1978 On the analysis of observational data: a discussion of the worth and uses of inductive techniques and respondent validation. Sociology 12(3): 545–552

Bloor M 1997 Techniques of validation in qualitative research: a critical commentary. In: Miller G, Dingwall R (eds) Context and method in qualitative research. Sage, London

Brink P 1991 Issues of reliability and validity. In: Morse J (ed) Qualitative nursing research: a contemporary dialogue. Sage, London, pp 164–186

Guba E, Lincoln Y 1985 Effective Evaluation: Improving the Usefulness of Evaluation. San Francisco, Jossey Bass

Guba E, Lincoln Y 1989 Fourth Generation Evaluation. Sage, California.

Hall J, Stevens P 1991 Rigor in feminist research. Advances in Nursing Science 13(3): 16–29

Hammersley M 1992 What's Wrong with Ethnography? Routledge, London

Hammersley M, Atkinson P 1995 Ethnography: Principles in Practice, (2nd edn). Routledge, London

Holloway I, Wheeler S 1996 Qualitative Research for Nurses. Blackwell Scientific, Oxford

Kirk J, Miller M 1986 Reliability and Validity in Qualitative Research. Sage, London
Koch T 1994 Establishing rigor in qualitative research: the decision trial. Journal of Advanced Nursing 19(5): 976–986
Lobiondo-Wood G, Haber J 1990 Nursing Research: Methods, Critical Appraisal and Utilisation, 2nd edn. Mosby, St Louis
Mason J 1996 Qualitative Researching. Sage, London
Myrdal G 1970 Objectivity in Social Research. Duckworth, London
Polit D, Hungler B 1989 Essentials of Nursing Research: Methods, Appraisal, and Utilization, 2nd edn. Lippincott, Philadelphia
Popper K 1979 Objective Knowledge: An Evolutionary Approach, 2nd edn. Clarendon Press, Oxford
Porter S 1993 Nursing research conventions: objectivity or obfuscation? Journal of Advanced Nursing 18(1): 137–143
Robson C 1993 Real World Research. Blackwell, Oxford
Sandelowski M 1986 The problem of rigor in feminist research. Advances in Nursing Science 8(3): 27–37
Schein E 1987 The Clinical Perspective in Fieldwork. Sage, London
Silverman D 1993 Interpreting Qualitative Data. Sage, London, pp 144–170

COMMENTARY

T. Long, M. Johnson Clinical Effectiveness in Nursing (2000) 4, 000–000 Rigour, reliability and validity in qualitative research

This is a fascinating and useful paper. On one level, it is a very valuable source of *definitions* of various terms and an examination of the differences between them. On another level, though, the authors may be a little too quick to reduce each of the 'alternatives' they find and to declare that they are simply a reworking of an old term. Thus, for the writers, 'dependability' is just another term for 'reliability' and 'credibility' for 'validity'. They may be right. But the reader is left waiting for the punchline: what are the authors going to offer in place of all this apparent playing with words?

Unfortunately, we are left waiting. The paper closes with the ambiguous sentence: 'The need is not for new criteria or novel terms but for difference means of addressing existing criteria.' Exactly what this means remains unclear. I would suspect that all the 'alternative' commentators that the authors quote would claim to be doing just that!

Personally, I feel that there really is a need for overall rigour in qualitative research. Because of the subjective nature of the qualitative project, there is a great danger that any form of sampling, data collection and analysis will be acceptable as long as you can *describe* what you have done. This leaves qualitative researchers in nursing and other disciplines open to the criticism levelled by Gournay and Ritter (1997) that much existing nursing research '… amounts to no more than anecdotal accounts of nurses' and patients' experiences'.

If qualitative research is do more than simply report individuals' accounts of their experience and if it is to attempt to draw together those accounts and thus theory-build, the processes that are employed in that work must be subject to rigour and to reliability and validity checks. Theories built on a few people's idiosyncratic accounts of what they thought or felt on a particular day are likely to be fairly flimsy ones. This paper does much to increase the debate around these important issues.

REFERENCE

Gournay K, Ritter S 1997 What future for research in mental health nursing? Journal of Psychiatric and Mental Health Nursing 4: 441–442

Philip Burnard

Author response

We are in complete agreement with the need for rigour in qualitative studies which is eloquently presented in this commentary. Such need is expressly declared in the opening to the article. Researchers of whatever philosophical persuasion, practitioners wishing to implement research findings, and consumers who will experience the consequences all may justifiably demand this.

The argument here is that semantic conjuring is not the solution. Alternative contributions commonly focus on an alleged need for different terms, declaring these to represent criteria of rigour which are distinct from traditional versions. Analysis of these claims to variation establishes that they are groundless. What is required is a raft of means to assess and demonstrate the degree of fulfilment of existing criteria which are appropriate and meaningful when applied to qualitative studies. The key (the 'punchline') is to focus on alternative means rather than alternative criteria.

The latter part of the article offers a contribution to the analysis of potential means to assess rigour. Brief critique is provided of some more commonly discussed strategies. Some are found to be less convincing than is often claimed (such as respondent validation). Others, audit of the decision trail for example, are recommended. The process displayed here is another example: open debate among contributors, peers and colleagues; and constructive criticism of method, findings and arguments can contribute to the required assessment. Whatever criteria or means are employed, their use must be based on sound evidence or reasoning, and the researcher prepared to answer to searching, constructive criticism.