Grading gems: Appraising the quality of research about social work and social care.

Taylor, B. J., Dempster, M., & Donnelly, M. (2007). Grading gems: Appraising the quality of research about social work and social care. DOI: 10.1093/bjsw/bch361

Published in:
British Journal of Social Work

Document Version:
Publisher's PDF, also known as Version of record

British Journal of Social Work (2007) **37,** 335–354
doi:10.1093/bjsw/bch361
Advance Access publication October 31, 2005

Grading Gems: Appraising the Quality of Research for Social Work and Social Care

Brian J. Taylor, Martin Dempster and Michael Donnelly

Brian Taylor is Lecturer in Social Work at the University of Ulster. Martin Dempster is Lecturer in Psychology, and Michael Donnelly is Reader in the Department of Epidemiology at Queen's University Belfast. *An earlier version of some of this material was presented at a conference: Taylor BJ and Donnelly M 'Graded grains make finer findings: Appraising the quality of a wide range of designs and methods', paper presented at the XII Cochrane Colloquium, Ottawa, Canada, 1–6 October 2004.*

Correspondence to Dr Brian Taylor, University of Ulster at Jordanstown, Shore Road, Newtownabbey BT37 0QB, Northern Ireland. E-mail: *bj.taylor@ulster.ac.uk*

Summary

The impetus towards basing practice and policy decisions more explicitly on sound research requires tools to facilitate the systematic appraisal of the quality of research encompassing a diverse range of methods and designs. Five exemplar tools were developed and assessed in terms of their usefulness in selecting studies for inclusion in a systematic review. The widely used 'hierarchy of evidence' was adapted and used to appraise internal validity. Four tools were then developed to appraise the external validity dimensions of generalizability (two scales) and methods of data collection (two scales). Methods of combining the scores generated by each tool were explored. Qualitative and quantitative studies were appraised, not separated into two spheres but by using complementary tools developed to appraise different aspects of rigour. There was a high level of agreement between researchers in applying the tools to twenty-two studies on decision making by professionals about the long-term care of older people. The scales for internal validity and generalizability discriminated between the qualities of studies appropriately. The two tools to appraise data collection gave diverse results. Excluding studies that scored in the lowest category on any scale appeared to be the scoring system that was most justifiable. This approach is presented to stimulate debate about the practical application of the evidence-based initiative to social work and social care. This study may assist in developing clearer definitions and common language about appraising rigour that should further the process of selecting robust research for synthesis to inform practice and policy decisions.

Keywords: appraisal of research, systematic review, evidence-based practice.

Context

There is an emerging consensus within health and social care that we should be striving for 'evidence-based practice', perhaps best defined as 'Conscientious, explicit and judicious use of the current best evidence in making decisions about the care of individual patients [and clients]' (Sackett *et al*., 1996, p. 71). Social work has always had formal and more informal research that has influenced policy and practice (Webb and Webb, 1932). In that sense, practice has always been 'evidence-based'. So what has changed that Trinder can claim 'the emergence of evidence-based practice has to be one of the success stories of the 1990s' (Trinder, 2000, p. 1)? Perhaps we are now questioning beliefs in methods that have not been tested (Bilsker and Goldner, 2000). Perhaps greater accountability is now required in public services (Davies *et al*., 2000). Perhaps there is a gap between the everyday work of practitioners and studies carried out by researchers (Seidl, 1991).

In one sense, the virtues of basing one's practice on the best available evidence are self-evident, as professionals (1) seek the best practice in a particular circumstance for the benefit of the client; (2) seek to make the best use of resources; and (3) seek decisions that are based on professional knowledge rather than dominated by organizational 'requirements' or pressures of political expediency (Macdonald and Sheldon, 1998; Sheldon, 2001; Webb, 2001). On the other hand, there are critics who see the present wave of enthusiasm as overly simplistic and constraining professional autonomy (Trinder, 2000), just as there have been critics in other professions (Dickersin and Berlin, 1992). There are those who prefer the term 'research-minded' practice (Thompson, 2000), thereby weakening the challenge to create sound research evidence to shape practice. Some see the technical aspects as being in conflict with value issues, although the social work role has always involved contributing professional knowledge and skills within a process that includes the values of the client, the profession and the society (Taylor and Devine, 1993). Some question the achievability of 'evidence based practice' (Ainsworth and Hansen, 2002) whilst others are identifying and addressing specific challenges (Sheldon, 1987; Sheldon and Chilvers, 2000; Pritchard, 2002*a*, 2002*b*; Webb, 2002). Similar debates have occurred in other professions (McAlister *et al*., 2000).

Interest in evidence-based practice certainly seems to be linked to the 'information age', although current concerns about the bewildering rate of growth of information sources available to practitioners (Needham, 2000) are reminiscent of Olive Stevenson's lament in this journal about 'the knowledge explosion' over thirty years ago (Stevenson, 1971). The accumulation of research findings makes the integration of knowledge into practice based on best evidence increasingly complex (Petitti, 1994), so more systematic approaches are required (Higgins and Pinkerton, 1998; Macdonald, 2001; Pawson *et al*., 2003). Attempts to measure effectiveness in social work are of course not new (French, 1952; Reid and Hanrahan, 1980; Sheldon, 1986; Macdonald and Sheldon, 1992). However, the impetus to have evidence of effectiveness has

renewed direction and energy supported by government policy (Department of Health, 1998; DHSSPS, 2001; Sanderson, 2000, 2002) and a range of initiatives in the UK (CRD, 2005; EPPI, 2005; NICE, 2005; SCIE, 2005) and internationally (Cochrane Collaboration, 2005; Campbell Collaboration, 2005).

It might now be regarded as not only a waste of resources but also unethical to undertake research if a study question has been answered adequately. Deciding the extent to which a research question has been addressed 'adequately' involves critically appraising the quality of research evidence so as to determine the strength of recommendations for practice and policy that might be based upon the findings (Grade Working Group, 2004). Systematic reviews of research (Dempster, 2003) might be regarded as embodying three key stages, the second of which is quality appraisal:

- searching for studies and identifying those that are relevant,
- appraising the quality of the identified studies, and
- extracting data from the selected studies and synthesis of the findings.

The need for models of critical appraisal of research in social work has been increasingly recognized but is not without conflicting views (Spittlehouse *et al.*, 2000).

Within the realms of quantitative research, particularly health care research, there is a 'Hierarchy of Evidence' (HoE) that has been developed over many years (Cochrane, 1973). This has achieved general acceptance as providing a basis for judging the rigour of a study of an intervention in terms of internal validity, which is the confidence that one can have in the results in terms of the cause and effect relationship being studied. The focus in the hierarchy is on quantitative studies. An example of this basic hierarchy for appraising research quality in health care (Khan *et al.*, 2004, reproduced in Figure 1) has three top sections that focus on: (1) fully experimental studies, such as randomized controlled trials; (2) quasi-experimental studies such as controlled trials; and (3) controlled observational research such as cohort studies. These key types of research design will be explained briefly to give a context to what follows.

Randomized controlled trials are regarded as the most rigorous design for demonstrating that an effect can be attributed to a particular cause (Chalmers

1. Experimental studies (e.g. randomised controlled trial with concealed allocation)
2. Quasi-experimental studies (e.g. experimental study without randomisation)
3. Controlled observational studies
   a. Cohort studies
   b. Case control studies
4. Observational studies without control groups
5. Expert opinion based on pathophysiology, bench research or consensus.

Figure 1 An example of the traditional hierarchy of evidence published by the Centre for Reviews and Dissemination, York

et al., 1981; Wortman 1994). Participants are randomly assigned to one of two or more groups, one group receiving the intervention being studied, and the other group(s) receiving 'normal treatment' or no treatment. Randomization is regarded as ensuring that any difference in outcome between groups is attributable to the intervention rather than to any other factor. An intervention might be a medical treatment, a type of psychosocial intervention more familiar in social work, or the way in which a service is organized, provided and delivered.

As a 'next-best' design, service users receiving a particular intervention might be matched with people with similar characteristics who do not receive the intervention. This is known as a 'case-controlled trial'. Although it lacks randomization, the matching of participants makes allowance for the influence of the characteristics chosen for matching.

Further down the hierarchy come studies where a group (cohort) of clients is measured before and after receiving an intervention, and where a control group that does not receive the intervention is measured at the same points in time. This provides some control for changes that might have occurred during the time period that are not due to the intervention. At the bottom of the hierarchy come studies without controls.

However, there are a number of criticisms of this hierarchy (Glasziou *et al.*, 2004), even though it is widely recognized as encapsulating key elements for appraisal of study quality (AHRQ, 2002). First, although it may be highly regarded for the appraisal of internal validity (causality), it is limited in addressing issues of external validity or relationship to the real world (Campbell and Stanley, 1966). Intervention studies that require a controlled environment, as free as possible from the impact of confounding or confusing factors, may be the most rigorous test of the effectiveness of a therapy, treatment or service. Yet, the appropriateness of applying this test to social care 'treatments' in a complex world characterized by poverty, stresses and multiple diseases is unclear. By their nature, measures to improve the internal validity of a particular study may reduce its external validity.

Second, the Hierarchy of Evidence is limited in addressing questions about the rigour of qualitative research. Qualitative studies tend to address questions such as 'why?' or 'how?' rather than 'what effect?' or 'how much effect?'. This may include data on social meanings, such as perceptions, opinions, experiences, interactions, feelings and views about health and social care processes and outcomes. Such data might be gathered through open questions in, for example, interviews, questionnaires and focus groups.

There are various approaches to appraising the quality of qualitative research (Drisko, 1997). Some (Burns, 1989; Barbour and Barbour, 2003; Pluye *et al.*, 2004) argue that a distinctly different set of criteria is required than for quantitative research, perhaps because of the variety of underpinning philosophical assumptions. The *British Medical Journal* has a range of checklists for different purposes, including one for 'Qualitative Research', but with the caveat that 'The BMJ's editors don't routinely use checklists for critical appraisal, but these are the kind of questions we ask ourselves when reading

papers' (BMJ, 2003). Globerman (1993) describes a training programme in critical appraisal skills for social workers, but the publication gives no system for ranking the research, or clear identification of domains of rigour.

Others argue that having separate criteria for qualitative and quantitative categories is not sufficient, but that separate criteria are required for each distinct research method or approach (Harden, 2004). As examples, Kirk and Miller (1986) focus on rigour in participant observation research, Muecke (1994) ethnographic studies and Corbin and Strauss (1990) grounded theory. The Critical Appraisal Skills Programme for health care (CASP, 2002) publishes separate lists for the appraisal of research using qualitative, cohort, case–control and randomized controlled trial designs, but does not differentiate between the various qualitative methods.

Another argument is that 'quality in qualitative research can be assessed with the same broad concepts of validity and relevance used for quantitative research, but these need to be operationalized differently' (Mays and Pope, 2000, p. 51). As examples, Elliott *et al.* (1999) and Walter *et al.* (2004) provide appraisal frameworks that include dimensions applicable to both qualitative and quantitative studies, although the former then go on to provide some additional dimensions that are regarded as relevant only to (all) qualitative studies.

The developments reported here examine the extent to which there is common ground between frameworks for the appraisal of qualitative and quantitative research. The purpose of the present study was to develop criteria for the appraisal of both quantitative and qualitative studies that would be applicable to both when operationalized appropriately. Accommodating the different philosophical assumptions underpinning research was addressed using an approach that was described in a recent report produced on behalf of the UK Cabinet Office (Spencer *et al.*, 2003, p. 50). A 'middle of the road' (rather than either extreme position) was adopted regarding the nature of knowledge and reality, the relationship between the researcher and the researched, and the relationship between facts and values.

Thus, the aim was to develop an approach that encompassed research into processes as well as studies of interventions, and that embraced a wider range of aspects of validity than the traditional Hierarchy of Evidence. Rather than seeking one hierarchy to cover all aspects, we sought to begin to develop a range of tools to appraise specific aspects of research design and methods.

In addition, a scoring system was developed so that the quality criteria might be used for a specific purpose such as deciding which studies to include in a review of research to guide practice. Various possibilities were explored, mindful of potential pitfalls but in the interest of stimulating debate and progress. The overall purpose was to facilitate the appraisal of research so as to further understanding about the validity of studies, to uncover reasons for differences among studies, and to provide information for readers to improve their judgment about the usefulness of studies for their work (cf. Meade and Richardson, 1997). Finally, an appraisal of the quality criteria was tested even though only a few studies have attempted to address this area (Sandelowski and Barroso, 2002 is an example).

The specific aims of this study were:

1 To investigate the extent to which the existing Hierarchy of Evidence could be modified and applied meaningfully to studies of social work and social care, and to the appraisal of the internal validity of a wider range of research designs than is covered by the present HoE tool.

2 To develop scales to appraise dimensions of external validity, and to evaluate these by applying them to a small number of studies that spanned a diverse range of methodologies.

3 To explore and test ways in which scales might be used to create criteria for inclusion in a systematic review.

Method

This study was undertaken as part of a systematic review of 'how professionals make decisions on the long-term care of older people'. The twenty-two studies considered here (Taylor *et al*., 2003) were identified as part of a systematic search of electronic databases (Taylor, 2003) followed by hand searching for articles published in the English language in journals with a blind peer review process. Only one paper reported an intervention study; the other papers described 'observational' studies (both descriptive and analytic) of decision processes, using a variety of quantitative and qualitative methods.

The intervention study (Kane *et al*., 1999) was a controlled before-and-after study of the impact of an intervention designed to increase case managers' responsiveness to the care-related values of their clients. The main elements of the intervention comprised a training programme, consensus development of a working procedure, and prompt cards given to older people. The twenty-one observational studies encompassed studies of the factors that are considered by decision makers (from social work, medicine and nursing); characteristics of professionals and organizations; and the impact of family, societal and resource issues on decision making. A variety of methodological approaches were employed by these studies, including:

– non-randomized controlled intervention study (two);
– before-and-after study with non-allocated interventions and controls (one);
– factorial survey (two);
– prospective longitudinal (cohort) study with concurrent controls (one);
– cross-sectional survey (twelve); and
– case study (four).

A key challenge for researchers, teachers, managers and practitioners is to find a parsimonious approach with which to appraise the quality of the diverse range of research that tends to characterize the field. Some dimensions of

quality do not lend themselves to quantification, even though they are a regular part of the quality appraisal processes used by journals. The appropriateness of the choice of design, data collection tool and method of analysis, and the effectiveness with which these are carried out, are important dimensions in appraising rigour (Mays and Pope, 2000; Giacomini and Cook, 2000*a*, 2000*b*). These dimensions appear to be addressed regularly during the peer review process of journals and are not considered here. The focus of the present study was on grading the quality of the methods used in published studies as an indication of the robustness of the results, accepting for the present purpose the appropriateness of the methods and the effectiveness of their implementation.

There were a number of issues relating to the merits or otherwise of using the traditional Hierarchy of Evidence to appraise internal validity. First, the Hierarchy of Evidence did not seem to give sufficient credence to the 'lower levels' in terms of differentiating and valuing appropriately the range of designs within this category. For example, survey studies with a comparator group are arguably more robust in terms of internal validity than studies without a comparator group. One study compared the decisions of front line social workers with the decisions of a multi-professional panel that included medicine, nursing and social work (Austin and Seidl, 1981). The methodological approach illustrated by this research (i.e. a study with a 'control' that did not use matched pairing of research participants) is an example of an approach that does not appear to be recognized and valued sufficiently by the traditional HoE.

Second, two studies were factorial surveys involving presenting decision makers with a unique realistic set of vignettes (case scenarios) with randomized characteristics based on factors found to be significant in previous studies (Taylor, in press). Health and social care professionals were asked to make decisions about clients presented in the vignettes (Hennessy, 1993; Degenholtz *et al*., 1999). The application of the HoE model appears to ignore the rigour of studies that employ factorial surveys by 'relegating' them to the level or category of 'observational studies'. Yet, it could be argued that these studies are similar to randomized controlled trials and with the added virtue of basing vignettes on real cases, thereby establishing good ecological validity.

Third, the language of 'trial' may not be meaningful for practitioners and policy makers in social work and social care. The term 'intervention study' appears to be a more apt and acceptable description of a study that examines the impact of a medical treatment or equally an intervention in social work, criminal justice or education.

The conclusion of these considerations was to propose an Internal Validity Scale, as illustrated in Figure 2.

Next, scales were created to appraise external validity. Two major dimensions were chosen: generalizability and data collection. This choice reflects major themes in the literature (Mays and Pope, 1995; Greenhalgh and Taylor, 1997; Mays and Pope, 2000; Peabody *et al*., 2000; Long and Godfrey, 2004). Two scales were created for each of the two dimensions (see Figures 3 and 4).

A Experimental designs
12 randomized controlled intervention study
11 non-randomized controlled intervention study
10 before-and-after study with non-allocated intervention and controls
09 interrupted time series intervention study without controls

B Survey designs with controls
08 Factorial survey
07 Prospective longitudinal survey ('cohort study') with concurrent controls
06 Retrospective longitudinal survey with concurrent controls (case–control)
05 Cross-sectional survey (or aggregating or comparing cases) with controls

C Survey designs without controls
04 Longitudinal survey with no controls
03 Cross-sectional survey (or aggregating or comparing cases) without controls
02 Case study (or a number of cases not aggregated or compared)
01 Expert opinion (including consensus methods)
NB 'Controls' may include a standard for comparison other than similar cases.

Figure 2 Internal Validity Scale

Sampling Method Scale
7 Complete census or total sample
6 Probability sampling from a defined sampling frame
5 Non-probability sampling from a defined frame (e.g. consecutive sampling)
4 Non-probability sampling from a non-defined sampling frame (e.g. snowball)
3 Convenience (non-systematic) sampling
2 Justifiable selection of single case or group of cases
1 Unjustifiable selection of single case or group of cases

Generalisability Appraisal Scale
6 Generalisable to similar professionals anywhere in the world
5 Generalisable to similar professionals in USA, EU or similar subcontinent
4 Generalisable to similar professionals in same nation, US state, or other entity with legal jurisdiction and control of public policy
3 Generalisable to similar professionals in same geographical region with definable culture, organisational arrangements, and minor legislative or public policy variation (e.g. Scotland, Wales, Northern Ireland, north of England, county, major conurbation, states of Australia, Canada etc.)
2 Generalisable to similar professionals in same organisation (local government authority, health *and* social services board area (NI), health *and* social services trust etc.)
1 Not clear to whom the study is generalisable.

Figure 3 Tools to appraise generalizability

One of the scales for generalizability, the Sampling Method Scale (within Figure 3), was based on a consideration of the method used to select a sample. Sampling was considered in relation to the main unit of analysis in a particular

Participant Task Realism Scale

4 Usual professional judgement or decision (or other work or life task)

3 Usual type of professional judgement or decision in normal setting (including vignettes concerned with decisions in present context, knowledge, skills, etc.)

2 Usual type of professional judgement or decision in abnormal setting (including interviews discussing recollections of or intentions regarding usual work or life tasks)

1 Unusual or inappropriate professional judgement or decision task.

Data Collection Impact Scale

4 Covert observation (covert direct observation, standardised actor, study of documents produced as part of normal work or life, etc.)

3 Overt contemporary data collection method (overt direct observation, vignettes or case scenarios, diaries, tests of contemporary facts, abilities, etc.)

2 Overt historical data collection method (focus groups, interviews, questionnaires, tests of past events, decisions, functioning, etc.)

1 Overt predictive data collection method ('what would you do in years to come if ...?')

Figure 4 Tools to appraise data collection

study, which might be a decision, a team or an organization as well as an individual person. The second scale for generalizability, the Generalizability Appraisal Scale (also within Figure 3), addressed the organizational, jurisdictional or geographical generalization that might reasonably be made for a study. Generalizability was considered in terms of logical generalizability as well as probabilistic generalizability. In other words, the standard of employing an appropriate sampling frame was appraised together with the degree to which, logically, one might generalize from the study. In practice, a key dimension was the extent to which the findings were generalizable from a particular organization in which a study was undertaken to wider organizational contexts.

The first scale to appraise data collection, the Participant Task Realism Scale (within Figure 4), considered the realism of the task that the participant was undertaking. The task of a research participant was considered to be an important aspect of ecological validity in terms of the extent to which performance on a set task or responses by participants were true-to-life as opposed to artificial.

The second scale for appraising data collection, the Data Collection Impact Scale (also within Figure 4), focused on the impact that a data collection tool might have in terms of biasing participants' responses. This scale considered data collection tools using categories ranging from covert observation through overt observation to historical and predictive scenarios (e.g. what would you do in years to come if . . . ?). In the process of attempting to assess the validity of data collection methods, a classification scheme (Figure 5) was developed to produce an agreed terminology, but this was not regarded as a hierarchy.

The five scales were applied to the twenty-two studies independently by two researchers (BJT and MDe). Any disagreements were discussed to reach

1 Observation
   a Subsequently verifiable observation with sight and sound (e.g. video-taped)
   b Subsequently verifiable observation with sound only (e.g. audio-taped)
   c Contemporaneously verifiable observation with sight and sound, e.g. one-way mirror with an observer
   d Contemporaneously verifiable observation with sound only (e.g. additional observer listening to telephone interview)
   e Un-verifiable observation or with only the observer's own written notes.
2 Standardized patient or client (actor)
   a Subsequently verifiable (taped) standardized patient or client (actor)
   b Unverifiable standardized patient or client (actor)
3 Documents
   a Documents relating to patients and clients
   b Other documents
4 Databases
   a Patient and client databases
   b Other databases
5 Vignettes (case scenarios)
   a Vignettes with randomized characteristics
   b Vignettes without randomized characteristics
6 Tests and Questionnaires (sub-codes (a) in person (b) by post (c) by telephone)
   a Tests and questionnaires that are validated and standardized
   b Tests and questionnaires that are validated
   c Tests and questionnaires that are standardized
   d Customised tests and questionnaires

NB Questionnaires by telephone regarded as strongly structured interviews (item 9 below)

7 Diaries
8 Group interviews or focus groups
   a Subsequently verifiable (taped) strongly structured group sessions
   b Subsequently verifiable (taped) semi-structured group sessions
   c Unverifiable strongly structured group sessions
   d Unverifiable semi-structured group sessions
9 Interviews with individuals (sub-codes (a) in person (b) by post (c) by telephone)
   a Subsequently verifiable (taped) strongly structured interviews
   b Subsequently verifiable (taped) semi-structured interviews
   c Unverifiable strongly structured interviews
   d Unverifiable semi-structured interviews

Figure 5 Classification of data collection methods

consensus, if necessary with a third researcher (MDo). The number of studies was too small to merit undertaking a statistical test of inter-rater reliability.

Once a score for each study on each of the five scales was determined, we sought to identify justifiable ways to use them to determine criteria for inclusion in the systematic review. The scoring on each scale was numbered from one indicating the lowest-quality category to simplify the process of exploring

meaningful ways in which scores from more than one scale could be combined, including:

1 Simple scoring (i.e. adding the total scores on the scales as given);

2 Weighted scoring (i.e. introducing a weighting factor, for example, to compensate for the varying length of scales, or to assign a degree of relative importance to each scale);

3 Setting a minimum criterion on each scale for inclusion, for example, to:

   a Eliminate any study in lowest category on any scale,

   b Include only those that score in the highest category (or categories), or

   c Set a minimum appropriate to that particular scale;

4 Combining the scores for the two generalizability scales, and similarly combining the scores from the data collection scales to give two measures for consideration alongside internal validity.

Results

The scores on each of the five scales for the twenty-two studies included in the review are shown in Table 1. The researchers readily achieved consensus in applying the tools to the range of methods used in the studies in the review.

In terms of generalizability, if studies that scored in the lowest category were excluded, then the Generalisability Appraisal Scale excludes seven studies, and the Sampling Method Scale two, both of which would be excluded by the Generalisability scale (Figure 3). The former seemed to be a useful criterion level for a broad-ranging review that would thereby include approximately two-thirds of the studies retrieved. On this occasion, the Sampling Method Scale might be regarded as redundant if the Generalisability Appraisal Scale were used. This congruence confirmed the compatibility of these scales, with one being more selective than the other. Requiring a combined score of four or more would have eliminated only four studies. In terms of the methods of the studies, the rationale for this was less clear than for the criterion above. Requiring a combined score of five or more would eliminate nine of the studies, including one that was relatively strong on data collection and internal validity criteria (Abrahams *et al*., 1989) and including one that scored only '1' on the Generalisability Scale (Kaufman, 1995). This seemed unreasonable and unbalanced for the body of data. Studies of pilot services tended to be eliminated by these more selective steps, although if the studies were well designed, they seemed worth including otherwise.

The two tools to appraise data collection (Figure 4) gave similar scores for some studies, but some studies scored higher on one scale and others higher on the other scale. As an example, a legal case study (Dubler, 1988) scored highly in that the data collection did not influence the respondents (Data Collection

Table 1 Appraisal scores for the studies in the review

Article	Internal Validity Scale (12H-1L)	External validity			
		Sampling Method Scale (7H -1L)	Generalisability Appraisal (6H-1L)	Participant Task Realism (4H–1L)	Data Collection Impact (4H–1L)
Abrahams, M. A., Capitman, J., Leutz, W. and Macko, P. (1989)	4	2	2	3	3
Austin, C. D. and Seidl, F. W. (1981)	5	6	3	4	4
Degenholtz, H., Kane, R., Kane, R. and Finch, M. D. (1999)	8	6	5	3	3
Dubler, N. N. (1988)	2	1	1	1	4
Gentry, J. W. and Kennedy, P. F., Macintosh, G. (1995)	3	6	3	2	1
Hennessy, C. H. (1987)	2	2	2	4	3
Hennessy, C. H. (1989)	2	2	2	4	3
Hennessy, C. H. (1993)	8	7	2	3	3
Hunter, S., Brace, S. and Buckley, G. (1993)	2	5	2	2	2
Kane, R. A., Degenholtz, H. B. and Kane, R. L. (1999)	10	7	4	4	2
Kaufman, S. R. (1995)	3	4	1	2	2
Lagergren, M. (1995)	3	7	3	4	4
Lagergren, M. and Johansson, P. A. (1998)	3	5	3	4	4
Mackay, R. and Lishman, J. (1991)	3	5	2	4	2
Mastrian, K. G. (2001)	3	3	1	2	3
Mastrian, K. G. and Dellasega, C. (1996)	3	3	1	2	2
McKeganey, N. P. (1991)	3	2	1	4	3
McKeganey, N., MacPherson, I. and Hunter, D. J. (1988)	3	2	1	4	3
Prager, E. (1986)	7	2	2	3	3
Schneider, R. L. and Kropf, N. P. (1996)	2	1	1	4	2
Seccombe, K., Ryan, R. and Austin, C. D. (1987)	3	6	4	2	2
Smith, B., O'Malley, S. and Lawson, J. (1993)	5	5	2	4	4

NB (1) Where more than one data collection method is used, the higher score is used. (2) Numbers in the column headers indicate the highest (H) and lowest (L) scores possible on that scale.

Impact Scale), but low in terms of the 'normality' of the matter under study (Participant Task Realism Scale). If studies that scored in the lowest category were excluded, then the Participant Task Realism Scale excludes this study and the Data Collection Impact Scale eliminates a different study (Gentry *et al.*, 1995, a marketing study using a data collection method rarely used in health and social services research).

Given the diverse scoring between the data collection scales, the possibility of a combined score was explored. Requiring a total score of four or more on data collection for inclusion would exclude only one of the studies excluded by the previous criterion (lowest score on any scale). Requiring a total score of five or more for inclusion eliminated four more studies (i.e. five in total), but, in general, this seemed unduly harsh given the short scales used for this. Interview studies (e.g. Seccombe *et al.*, 1987) and an anthropological study with some merit (Kaufman, 1995) would be among those excluded if this criterion were used. In general, the exclusion of those in the lowest category on either of these short data collection scales seemed the most robust use.

On the Internal Validity Scale (Figure 2), none scored in the lowest category. However, with this scale, it seemed more justifiable to 'draw the line' at a variety of points than with the external validity scales. If the only studies included were those that scored four or higher, then this would give seven studies with relatively high internal validity; including only those that scored three or higher would include seventeen studies. About a third of those included by this latter criterion would be excluded by even the most generous generalizability criterion discussed above. In general, the Internal Validity Scale seemed useable in a wider variety of ways than those created for external validity.

Discussion

The impetus towards basing practice and policy on robust evidence presents many challenges, not least in appraising the quality of research so as to be confident of its rigour and appropriateness for purpose. This study explored a variety of approaches to appraising generalizability and data collection as dimensions of external validity as well as developing the traditional hierarchy of evidence for appraisal of internal validity. Overall, the tools that were developed were useful in facilitating structured discussion and appraisal of the relative quality of studies according to a variety of key dimensions.

In terms of generalizability, the Sampling Method Scale and the Generalisability Appraisal Scale (Figure 3) seemed congruent with each other, with the latter being more selective than the former. The more broadly conceptualized generalizability scale seemed more useful than a simple appraisal of the method used to select the sample.

The data collection appraisal scales (Figure 4) gave similar results in the scores for many articles, but with some marked differences. Traditional tools of

qualitative research, such as focus groups and interviews, which rely on retrospective recounting of experiences, did not score highly using the Data Collection Impact Scale. However, such methods scored more highly on the Participant Task Realism Scale, although this scale is perhaps influenced particularly by the focus of this review on decision making. The data collection scales might have been more effective in discriminating between studies if there had been any publications with more artificial experimental conditions. Further exploration using different study questions is required.

In terms of scoring systems for deciding on inclusion in a review, simple addition of scores seemed inappropriate. Weighting would be an improvement, but depends on a decision as to the relative importance of the scales, even if it is to give them equal weighting by compensating for their differing lengths. The external validity tools did not seem appropriate for the task of including only those at the 'top' of the list, although the Internal Validity Scale seemed comfortable to use setting a criterion for inclusion at a variety of levels. The approach that seemed to make the most sense intuitively was to use the external validity tools to exclude those that fell into the lowest category on that scale. This would leave a wide range of research to be included, whilst excluding that which seemed less robust for the purpose.

It is recognized that the development of such tools will be contentious (Murphy *et al.*, 1998). However, some initiative is required if social work is to progress towards basing its practice explicitly on good-quality research. We may fail to recognize research 'gems' for their true worth unless there is some consensus on identifying and grading quality. The credibility of a synthesis of previous research is jeopardized if studies of questionable rigour are included. The approach described here proved effective in creating objective criteria for inclusion that accorded with the general principles of good research and the perceptions of the researchers. Although the evaluation is necessarily subjective, the study team had the benefit of a wide range of experience of qualitative and quantitative research methods applied in health and social services contexts. Appraising the tools on a body of literature using diverse methods but addressing a common research area was an advantage in clarifying thinking on the types of results obtained by each study and the possible bias inherent in each.

The general proposal arising from this work is that a more inclusive hierarchy of evidence should retain its key role in relation to research design in order to discharge the 'burden of proof' regarding causality; however, other aspects of method must also be appraised, with more attention to external validity. Although this presents the dilemmas inherent in combining the scores, it is argued here that this will be achieved most effectively by using one or more additional scales that might be used in different ways. The scales developed here illustrate possibilities, but are recognized as only one step forward in a lengthy process of development.

Finally, there are a number of benefits of the approach explored here. First, it will assist in the development of a common language around defined domains of rigour which should further the process of constructive dialogue

between advocates of different research methods. Second, this approach does not separate qualitative and quantitative studies into two totally separate spheres. Rather, it recognizes a 'round table' of designs and methods, and thus assists in bridging the gap between these complementary research approaches. Third, it provides a practical way forward to address an urgent need as part of the task of synthesizing robust research to inform practice and policy decisions for the ultimate benefit of clients and patients.

Accepted: July 2005

References

Abrahams, M. A., Capitman, J., Leutz, W. and Macko, P. (1989) 'Variations in care planning practice in the social/HMO: An exploratory study', *The Gerontologist*, **29**(6), pp. 725–36.

AHRQ (2002) *Systems to Rate the Strength of Scientific Evidence*, Summary Evidence Report/Technology Assessment No.47, US Department of Health and Human Services: Agency for Healthcare Research and Quality, available online at: www.ahrq.gov.

Ainsworth, F. and Hansen, P. (2002) 'Evidence-based social work: A reachable goal?', *Social Work and Social Services Review*, **10**(2), pp. 35–48.

Austin, C. D. and Seidl, F. W. (1981) 'Validating professional judgement in a home care agency', *Health and Social Work*, **6**(1), pp. 50–6.

Barbour, R. S. and Barbour, M. (2003) 'Evaluating and synthesizing qualitative research: The need to develop a distinctive approach', *Journal of Evaluation in Clinical Practice*, **9**(2), pp. 179–86.

Bilsker, D. and Goldner, E. (2000) 'Teaching evidence-based practice in mental health', *Research on Social Work Practice*, **10**(5), 664–9.

BMJ (British Medical Journal) (2003) Editors' Checklists: Critical Appraisal Questions, available online at: http://bmj.com/advice/checklists.shtml (accessed 1 November 2004).

Burns, N. (1989) 'Standards for qualitative research',*Nursing Science Quarterly*, **2**(1), pp. 44–52.

Campbell, D. T. and Stanley, J. C. (1966) *Experimental and Quasi-experimental Designs for Research*, Chicago, Rand McNally.

Campbell Collaboration (2005) (for reviews of the effectiveness of interventions in social welfare, education and criminal justice), available online at: http://www.campbellcollaboration.org (accessed 21 June 2005).

CASP (Critical Appraisal Skills Programme) (2002) *Appraisal Tools*, available online at:http://www.phru.nhs.uk/casp/appraisa.htm (accessed 1 November 2004).

Chalmers, T. C., Smith, H. J. R., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Ambroz, A. (1981) 'A method for assessing the quality of a randomised control trial', *Controlled Clinical Trials*, **2**(1), pp. 31–49.

Cochrane, A. L. (1973) *Effectiveness and Efficiency: Random Reflections on Health Services*, London, Nuffield Provincial Hospitals Trust.

Cochrane Collaboration (2005) (for reviews of the effectiveness of interventions in health and social care), available online at: http://www.cochrane.org (accessed 21 June 2005).

Corbin, J. and Strauss, A. (1990) 'Grounded theory research: Procedures, canons, and evaluation criteria', *Qualitative Sociology*, **13**(1), pp. 3–21.

CRD (National Health Service Centre for Reviews and Dissemination) (2005), available online at: http://www.york.ac.uk/inst/crd (accessed 21 June 2005).

Davies, H. T. O., Nutley, S. M. and Smith, P. C. (2000) *What Works? Evidence-Based Policy and Practice in Public Services*, Bristol, The Policy Press.

Degenholtz, H., Kane, R., Kane, R. and Finch, M. D. (1999) 'Long-term care case managers' out-of-home placement decisions: An application of hierarchical logistic regression', *Research in Aging*, **21**(2), pp. 240–74.

Dempster, M. (2003) 'Systematic review', in Miller, R. and Brewer, J. (eds), *The A to Z of Social Research*, London, Sage.

Department of Health (1998) *Modernising Social Services*, Cm 4169, London, The Stationery Office.

DHSSPS (2001) *Best Practice—Best Care: A Framework for Setting Standards, Delivering Services and Improving Monitoring and Regulation in the HPSS: A Consultation Paper*, Belfast, Department of Health, Social Services and Public Safety.

Dickersin, K. and Berlin, J. A. (1992) 'Meta-analysis: State-of-the-science', *Epidemiologic Reviews*, **14**, pp. 154–76.

Drisko, J. W. (1997) 'Strengthening qualitative studies and reports: Standards to promote academic integrity', *Journal of Social Work Education*, **33**(1), pp. 185–97.

Dubler, N. N. (1988) 'Improving the discharge planning process: Distinguishing between coercion and choice', *The Gerontologist*, **28**(Suppl.), pp. 76–81.

Elliott, R., Fischer, C. T. and Rennie, D. L. (1999) 'Evolving guidelines for publication of qualitative research studies in psychology and related fields', *British Journal of Clinical Psychology*, **38**(3), pp. 215–29.

EPPI (Evidence for Policy and Practice Information and Co-ordinating Centre) (2005), available online at: http://eppi.ioe.ac.uk/EPPIWeb (accessed 21 June 2005).

French, D. G. (1952) *An Approach to Measuring Results in Social Work*, New York, Columbia University Press.

Gentry, J. W., Kennedy, P. F. and Macintosh, G. (1995) 'Marketing implications of the expected role of physicians in family decisions concerning the institutionalization of the elderly', *Psychology and Marketing*, **12**(7), pp. 647–62.

Giacomini, M. K. and Cook, D. J. (2000*a*) 'Users' guides to the medical literature: XXIII. Qualitative research in health care A—Are the results of the study valid?', *Journal of the American Medical Association*, **284**(3), p. 357.

Giacomini, M. K. and Cook, D. J. (2000*b*) 'Users' guides to the medical literature: XXIII. Qualitative research in health care B—What are the results and how do they help me care for my patients?', *Journal of the American Medical Association*, **284**(4), p. 478.

Glasziou, P., Vandenbroucke, J. and Chalmers, I. (2004) 'Assessing the quality of research', *British Medical Journal*, **328**(7430), pp. 39–41.

Globerman, J. (1993) 'Teaching critical appraisal of the social work literature', *Journal of Teaching in Social Work*, **7**(2), pp. 63–79.

Grade Working Group (2004) 'Grading quality of evidence and strength of recommendations', *British Medical Journal*, **328**, 1490.

Greenhalgh, T. and Taylor, R. (1997) 'How to read a paper: Papers that go beyond numbers (qualitative research)', *British Medical Journal*, **315**(7109), pp. 740–3.

Harden, A. (2004) 'A review of tools for assessing the quality of qualitative studies: Implications for systematic reviews', paper presented at the XII Cochrane Colloquium, Ottawa, Canada, 1–6 October 2004.

Hennessy, C. H. (1987) 'Risks and resources: Service allocation decisions in a consolidated model of long term-care', *Journal of Applied Gerontology*, **6**(2), pp. 139–55.

Hennessy, C. H. (1989) 'Autonomy and risk: The role of client wishes in community-based long-term care', *The Gerontologist*, **29**(5), pp. 633–9.
Hennessy, C. H. (1993) 'Modeling case management decision-making in a consolidated long-term care program', *The Gerontologist*, **33**(3), pp. 333–41.
Higgins, K. and Pinkerton, J. (1998) 'Literature reviewing: Towards a more rigorous approach', in Iwaniek, D. and Pinkerton, J. (eds), *Making Research Work: Promoting Child Care Policy and Practice*, Chichester, Wiley.
Hunter, S., Brace, S. and Buckley, G. (1993) 'The inter-disciplinary assessment of older people at entry into long-term institutional care: Lessons for the new community care arrangements', *Research, Policy and Planning*, **11**(1 and 2), pp. 2–9.
Kane, R. A., Degenholtz, H. B. and Kane, R. L. (1999) 'Adding values: An experiment in systematic attention to values and preferences of community long-term care clients', *Journal of Gerontology: Social Sciences*, **54b**(2), pp. S109–19.
Kaufman, S. R. (1995) 'Decision making, responsibility, and advocacy in geriatric medicine: Physician dilemmas with elderly in the community', *The Gerontologist*, **35**(4), pp. 481–8.
Khan, K. S., ter Riet, G., Popay, J., Nixon, J. and Kleijnen, J. (2004) 'Stage II Conducting the Review—Phase 5: Study Quality Assessment', in Centre for Reviews and Dissemination, *Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for those Carrying Out or Commissioning Reviews Report Number 4 (2nd Edition)*, York, CRD.
Kirk, J. and Miller, M. C. (1986) *Reliability and Validity in Qualitative Research*, London, Sage Publications.
Lagergren, M. (1995) 'Determining the appropriate level of care', *Scandinavian Journal of Social Medicine*, **23**(3), pp. 209–15.
Lagergren, M. and Johansson, P. A. (1998) 'Are there differences in standard of care for the elderly? A comparative study of assistance decisions in Stockholm', *Scandinavian Journal of Social Welfare*, **7**(4), pp. 340–9.
Long, A. and Godfrey, M. (2004) 'An evaluation tool to assess the quality of qualitative research studies', *International Journal of Social Research Methodology*, **7**(2), pp. 181–96.
Macdonald, G. (2001) *Effective Interventions for Child Abuse and Neglect: An Evidence-based Approach to Planning and Evaluating Interventions*, Chichester, John Wiley and Sons Ltd.
Macdonald, G. M. and Sheldon, B. (1992) 'Contemporary studies of the effectiveness of social work', *British Journal of Social Work*, **22**(6), pp. 615–43.
MacDonald, G. and Sheldon, B. (1998) 'Changing one's mind: The final frontier?', *Issues in Social Work Education*, **18**(1), pp. 3–25.
Mackay, R. and Lishman, J. (1991) 'Assessment for residential care for old people in the north of Scotland', *Practice*, **5**(4), pp. 249–64.
Mastrian, K. G. (2001) 'Differing perceptions in defining safe independent living for elders', *Nursing Outlook*, **49**(5), pp. 231–7.
Mastrian, K. G. and Dellasega, C. (1996) 'Helping families with long-term care decisions', *Caring Magazine*, **15**(2), pp. 68–9 and 71–2.
Mays, N. and Pope, C. (1995) 'Qualitative research: Rigour and qualitative research', *British Medical Journal*, **311**, pp. 109–12.
Mays, N. and Pope, C. (2000) 'Assessing quality in qualitative research', *British Medical Journal*, **320**, pp. 50–2.
McAlister, F. A., Staus, S. E., Guyatt, G. H. and Haynes, R. B. (2000) 'Users' guides to the medical literature: XX. Integrating research evidence with the care of the individual patient', *Journal of the American Medical Association*, **283**(21), pp. 2829–36.

McKeganey, N., MacPherson, I. and Hunter, D. J. (1988) 'How "they" decide: Exploring professional decision-making', *Research, Policy and Planning*, **6**(1), pp. 15–19.

McKeganey, N. P. (1991) 'The role of clinicians in residential home assessments', *Social Policy and Administration*, **25**(2), pp. 149–59.

Meade, M. O. and Richardson, W. S. (1997) 'Selecting and appraising studies for a systematic review', *Annals of Internal Medicine*, **127**(7), pp. 531–7.

Muecke, M. (1994) 'On the evaluation of ethnographies', in Morse J. M. (ed.), *Critical Issues in Qualitative Research Methods*, Thousand Oaks, CA, Sage Publications.

Murphy, E., Dingwall, R., Greatbatch, D., Parker, S. and Watson, P. (1998) 'Qualitative research methods in health technology assessment: A review of the literature', *Health Technology Assessment*, **2**(16).

Needham, G. (2000) 'Research and practice: Making a difference', in Gomm, R. and Davies, C. (eds), *Using Evidence in Health and Social Care*, London, Sage Publications and The Open University.

NICE (National Institute for Clinical Excellence) (2005), available online at: http://www.nice.org.uk (accessed 21 June 2005).

Pawson, R., Boaz, A., Grayson, L., Long, A. and Barnes, C. (2003) *Knowledge Review No 3: Types and Quality of Knowledge in Social Care*, London, Social Care Institute for Excellence.

Peabody, J. W., Locke, J., Glassman, P., Dresselhaus, T. R. and Lee, M. (2000) 'Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of 3 methods for measuring quality', *Journal of the American Medical Association*, **283**(13), pp. 1715–22.

Petitti, D. B. (1994) *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis*, Oxford, Oxford University Press.

Pluye, P., Grad, R., Dunikowski, L. and Stephenson, R. (2004) 'A challenging mixed literature review experience', paper presented at the XII Cochrane Colloquium, Ottawa, Canada, 1–6 October 2004.

Prager, E. (1986) 'Bureaucracy's impact on decision making in long-term care', *Health and Social Work*, **11**(4), pp. 275–85.

Pritchard, C. (2002*a*) 'The extremes of child abuse: A macro approach to measuring effective prevention', in Smith, D. (ed.), *Social Work and Evidence-Based Practice*, London and Philadelphia, Jessica Kingsley.

Pritchard, C. (2002*b*) 'Effective social work: A micro approach—reducing truancy, delinquency and school exclusions', in Smith, D. (ed.), *Social Work and Evidence-Based Practice*, London and Philadelphia, Jessica Kingsley.

Reid, W. J. and Hanrahan, P. (1980) 'The effectiveness of social work: Recent evidence', in Goldberg, E. M. and Connelly, N. (eds), *Evaluative Research in Social Care*, London, Heineman.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B. and Richardson, W. S. (1996) 'Evidence based medicine: What it is and what it isn't', *British Medical Journal*, **312**(7023), pp. 71–2.

Sandelowski, M. and Barroso, J. (2002) 'Reading qualitative studies', *International Journal of Qualitative Method*, **1**(1), article 5.

Sanderson, I. (2000) 'Evaluation in complex policy systems', *Evaluation*, **6**(4), pp. 433–54.

Sanderson, I. (2002) 'Evaluation policy learning and evidence-based policy making', *Public Administration*, **80**(1), pp. 1–22.

Schneider, R. L. and Kropf, N. P. (1996) 'The admission process in nursing homes: A clinical model for ethical decision-making', *Journal of Long-Term Home Health Care*, **15**(3), pp. 39–46.

SCIE (Social Care Institute for Excellence) (2005), available online at: http://www.scie.org.uk (accessed 21 June 2005).

Seccombe, K., Ryan, R. and Austin, C. D. (1987) 'Care planning: Case managers' assessment of elders' welfare and caregivers' capacity', *Family Relations*, **36**(2), pp. 171–5.

Seidl, F. W. (1991) 'Foreword', in Gibbs, L. E. (ed.), *Scientific Reasoning for Social Workers: Bridging the Gap Between Research and Practice*, New York, Macmillan.

Sheldon, B. (1986) 'Social work effectiveness experiments: Review and implications', *British Journal of Social Work*, **16**(2), pp. 223–42.

Sheldon, B. (1987) 'Implementing findings from social work effectiveness research', *British Journal of Social Work*, **17**(6), pp. 573–86.

Sheldon, B. (2001) 'The validity of Evidence Based Practice in Social Work: A reply to Stephen Webb', *British Journal of Social Work*, **31**, pp. 801–09.

Sheldon, B. and Chilvers, R. (2000) *Evidence Based Social Care: A Study of Prospects and Problems*, Lyme Regis, Russell House Publishing.

Smith, B., O'Malley, S. and Lawson, J. (1993) 'The costs and experiences of caring for sick and disabled geriatric patients: Australian observations', *Australian Journal of Public Health*, **17**(2), pp. 131–4.

Spencer, L., Ritchie, J., Lewis, J. and Dillon, L. (2003) *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*, London, Cabinet Office.

Spittlehouse, C., Acton, M. and Enock, K. (2000) 'Introducing critical appraisal skills training in UK social services: Another link between health and social care', *Journal of Interprofessional Care*, **14**(4), pp. 397–404.

Stevenson, O. (1971) 'Knowledge for social work', *British Journal of Social Work*, **1**(2), pp. 225–37.

Taylor, B. J. (2003) 'Literature searching', in Miller, R. and Brewer, J. (eds), *The A to Z of Social Research*, London, Sage.

Taylor, B. J. (in press) 'Factorial surveys: Using vignettes to study professional judgement', *British Journal of Social Work*.

Taylor, B. J. and Devine, T. (1993) *Assessing Needs and Planning Care in Social Work*, Hampshire, Ashgate.

Taylor, B. J., Dempster, M. and Donnelly, M. (2003) 'Hidden gems: Systematically searching electronic databases for research publications for social work and social care', *British Journal of Social Work*,**33**(4), pp. 423–39.

Thompson, N. (2000) *Theory and Practice in Human Services*, Buckingham, Open University Press.

Trinder, L. (2000) 'Evidence-based practice in social work and probation', in Trinder, L. and Reynolds, S. (eds), *Evidence-Based Practice: A Critical Appraisal*, Oxford, Blackwell Science.

Walter, I., Nutley, S., Percy-Smith, J., McNeish, D. and Frost, S. (2004) *Knowledge Review 7: Improving the Use of Research in Social Care Practice*, London, Social Care Institute for Excellence.

Webb, S. (2002) 'Evidence based practice and decision analysis in social work: An implementation model', *Journal of Social Work*, **2**(1), pp. 45–63.

Webb, S. and Webb, B. (1932) *Methods of Social Study*, Cambridge, London School of Economics and Political Science and Cambridge University Press.

Webb, S. A. (2001) 'Some considerations on the validity of evidence-based practice in social work', *British Journal of Social Work*, **31**, pp. 57–79.

Wortman, P. M. (1994) 'Judging research quality', in Cooper, H. and Hedges, L. V. (eds), *The Handbook of Research*, New York, Russell Sage Foundation.

Acknowledgements

The first author was supported in the research of which this formed part, and its dissemination, by Fellowships from the Research and Development Office of the Department of Health, Social Services and Public Safety for Northern Ireland, and the Economic and Social Research Council, London. This support, and that of Causeway Health and Social Services Trust, is gratefully acknowledged. Views expressed are those of the authors, and do not necessarily reflect the views of the above bodies.