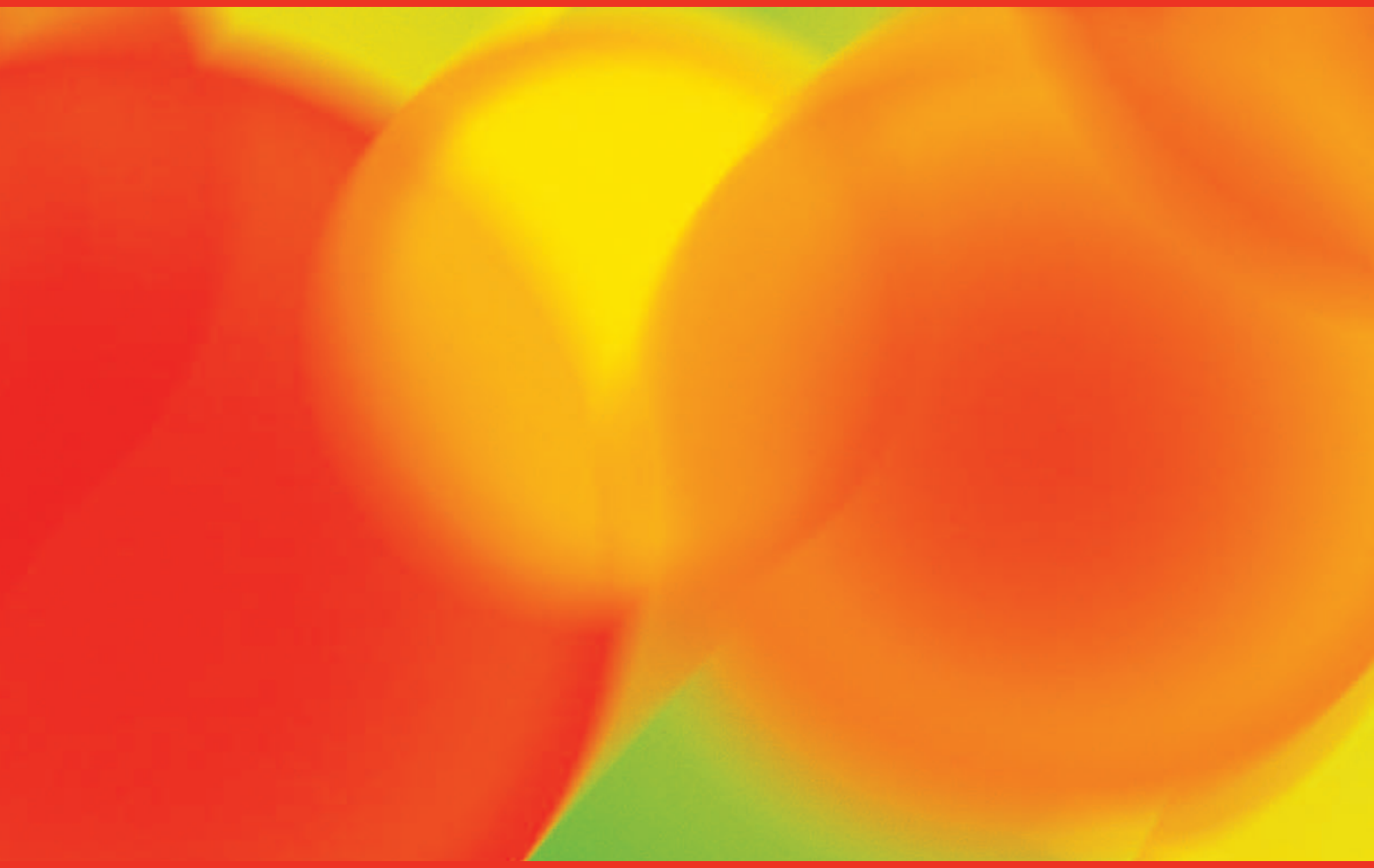
Evaluation and Evidence, Discussion Paper 1



Evaluating Outcomes in Social Work Education

January 2005

social care
institute for excellence

SCOTTISH INSTITUTE
FOR EXCELLENCE IN
SOCIAL WORK EDUCATION

First published in the UK by the Scottish Institute for Excellence in Social Work Education (SIESWE) and the Social Care Institute for Excellence (SCIE) in January 2005.

Scottish Institute for Excellence in Social Work Education
University of Dundee
Gardyne Road Campus
Dundee DD1 5NY
www.sieswe.org

Social Care Institute for Excellence
1st Floor
Goldings House
2 Hay's Lane
London SE1 2HB
www.scie.org.uk

British Library Cataloguing in Publication Data

A catalogue record for this publication is available from the British Library - ISBN 0-9549544-0-8

Dr John Carpenter is Professor of Social Work at Durham University.

Evaluating Outcomes in Social Work Education

Evaluation and Evidence, Discussion Paper 1

John Carpenter
School of Applied Social Sciences
University of Durham

social care
institute for excellence

SCOTTISH INSTITUTE
FOR EXCELLENCE IN
SOCIAL WORK EDUCATION

Contents

Preface

We live in an age where efficacy is a key issue, particularly in respect of the use of public funds. A poor evidence base underpinning policy or practice is, therefore, a vulnerability that should be avoided. Our two organisations have been concerned for some time to promote and support action within education for social work and social care that raises awareness of the importance of evidence-based practice and demonstrates a practical commitment to evaluating teaching and learning processes. We commissioned Professor Carpenter to write this discussion paper as a contribution to this process and were pleased at the positive reponse he received when introducing it at the 2004 Joint Social Work Education Conference in Glasgow. He brings to this task a long-standing interest and experience in educational evaluation, most recently in post-qualifying mental health programmes for the NHS Executive.

The paper addresses key aspects of the evaluation of changes in knowledge, skills and behaviour that can be attributed to, or expected to result from, learning processes within programmes. As such, the focus is different from, but complementary to, the evaluation of the impact of whole programmes per se, e.g. the introduction of the new social work degrees across the U.K.

The literature search carried out for this paper, and related studies previously commissioned by SCIE, have highlighted the paucity of reliable studies on the effectiveness of educational processes in this field. Similarly, there is almost no current body of knowledge examining the impact of training on subsequent practice in social care settings. Evaluation of teaching in both campus and workplace settings is regularly carried out using standard learner-feedback methods (normally questionnaires) but, as the author points out, this tells us little about how effective the learning has been. Data is also routinely collected on learners at the outset of modules and programmes but this is rarely used to establish baselines against which improvements in skills and knowledge can be measured.

In publishing this paper and supporting activities that follow on from it, our hope is that both SIESWE and SCIE can assist in remedying these deficiencies. We are joining forces to develop practical initiatives to assist social work educators to respond to the challenges of evaluating teaching and learning and to raise awareness among funding bodies of the need to support this type of research.

Professor Bryan Williams
Scottish Institute for Excellence in Social Work Education (SIESWE)

Professor Mike Fisher
Social Care Institute for Excellence (SCIE)

January 2005

1. Introduction

Recent systematic reviews to underpin social work education (Crisp et al., 2003; Trevithick et al., 2004) have identified the scarcity of evaluative research on the outcomes of methods of social work education; narrative accounts are plentiful, but it is rare to encounter an evaluation with carefully designed outcomes, and even more rare to find a controlled evaluation. For example, the knowledge review of assessment in social work identified 60 papers which described training programmes, but only 11 of these reported any information about their impact (Crisp et al. 2003, p. 35) and only one of these was (non-randomly) controlled. Similarly, the review of communication skills concluded that:

The review highlighted a dearth of writing which addressed the challenging issues of evaluating the learning and teaching of communication skills. This situation has serious implications for the issues of transferability…as without robust evaluative strategies and studies the risks of fragmented and context-restricted learning are heightened. (Trevithick et al., 2004, p.28)

This problem is not unique to social work. For example, Hullsman et al. (1999) reviewed the literature on teaching communication skills to clinically experienced doctors. They found only 14 studies published in the previous 13 years, most of which used "inadequate" research designs. However, these studies generally included multiple methods of assessing outcomes, for example behavioural observations and attempts to assess benefits for patients in addition to the learners' self-ratings of their skills. Unfortunately, they concluded that "studies with the most adequate designs report the fewest positive training effects".

Similarly, Bailey et al. (2003) surveyed all mental health trusts and social services departments in England asking about the evaluation of postqualifying training. Only 26 of the 66 organisations which responded (response rate 25%) said that they systematically evaluated training initiatives and in almost all cases the evaluation was confined to the trainees' satisfaction with the programmes provided.

The poor quality of research design of many studies, together with the limited information provided in the published accounts are major problems in establishing an evidence base for social work education. A systematic review of interprofessional education for the Learning and Teaching Support Network in Health Sciences and Practice (Freeth et al., 2002) initially identified 217 evaluation studies. However, after review three quarters were deemed inadequate (p.19), leaving only 53 studies of adequate quality. Half of these employed simple before-and-after designs with no comparison or control group and are therefore unable to provide a convincing demonstration of cause and effect (Freeth et al., 2002, p.54).

It seems fair to conclude with all the above reviewers that more and better quality evaluations need to take place. But if this is to happen, we need first of all to be clear about what we are trying to evaluate and then consider how this might be done. The aim of this paper is to stimulate discussion amongst educators and evaluators by attempting:

1. To identify what we mean by the 'outcomes' of social work education
2. To consider how these outcomes might be measured
3. To assess the advantages and disadvantages of different research designs for the evaluation of outcomes in social work education
4. To illustrate some of the methods and measures which have been used to evaluate outcomes

In order to achieve these ends, I will refer first to a review of theories of learning outcome by Kraiger et al., (1993) in order to clarify the conceptual basis for a discussion of learning outcomes. This review provides a synthesis of research in education and training which elucidates the relationships between factors which are thought to influence outcomes. I will draw on reviews of outcome studies from nursing and medicine as well as social work in order to illustrate what has been achieved so far.

The poor quality of research design of many studies, together with the limited information provided in the published accounts are major problems in establishing an evidence base for social work education.

The emphasis of this paper is on outcomes and how they might be evaluated. I will not therefore be concerned here with philosophies of education, curriculum design or the desirability or otherwise of particular modes of learning or course content. Similarly, I will not consider how we might research the *process* of social work education, except to mention one standardised observational approach by which we may describe the methods of teaching used by educators. This is not because I think that an understanding of process is unimportant; information about the mechanisms by which outcomes are produced and the context in which this happens is essential to evaluation research (Pawson and Tilley, 1997). These aspects of a programme should always be included in an evaluation report.

I should stress that the focus here is on the evaluation of programme level methods of teaching and learning rather than the global outcomes assessed by Marsh and Triseliotis (1996) and Lyons and Manion (2004) who examined newly qualified social workers' "readiness to practice", or the fit between social work education and agency expectations (Marsh and Triseliotis, 1996, p.2). Similarly, the focus of this paper is complementary to the approach being taken in research commissioned by the Department of Health in England on the evaluation of the new social work degree; that project is concerned with the characteristics, motivations, expectations and experiences of the students and, as a key outcome, degree completion rates. In contrast, the outcomes I will be considering here are more specific and detailed and relate to changes in knowledge, attitudes and behaviour which may be attributed to teaching and learning opportunities.

2. What Do We Mean By Outcomes?

The best known and most widely used classification of educational outcomes was devised by Kirkpatrick (1967). This model defined four levels of outcomes: learners' reactions to the educational experience; learning, conceptualised mainly as the acquisition of knowledge and skills; behaviour change, including the application of learning to the work setting; and results, assessed in relation to intended outcomes. This model was elaborated by Barr et al. (2000) for a review of interprofessional education in order to include the modification of attitudes as a learning outcomes and to divide "results" into change in organisational practice and benefits to patients/clients. The Kirkpatrick/Barr model was used by Freeth et al. (2002) to classify studies in a review of interprofessional education and by Bailey et al. (2003) for a review of postqualifying education in mental health. A generalised version is shown in Table 1.

Table 1. Levels of Outcomes of Educational Programmes (After Kirkpatrick, 1967 and Barr et al., 2000)

Level 1: ***Learners' Reaction*** – These outcomes relate to the participants' views of their learning experience and satisfaction with the training.

Level 2a: ***Modification in Attitudes and Perceptions*** – Outcomes here relate to changes in attitudes or perceptions towards service users and carers, their problems and needs, circumstances, care and treatment.

Level 2b: ***Acquisition of Knowledge and Skills*** – This relates to the concepts, procedures and principles of working with service users and carers. For skills this relates to the acquisition of thinking/problem solving, assessment and intervention skills.

Level 3: ***Changes in Behaviour*** - This level covers the implementation of learning from an educational programme in the workplace, prompted by modifications in attitudes or perceptions, or the application of newly acquired knowledge and skills.

Level 4a: ***Changes in Organisational Practice*** – This relates to wider changes in the organisation/delivery of care, attributable to an education programme.

Level 4b: ***Benefits to Users and Carers*** – This final level covers any improvements in the well-being and quality of life of people who are using services, and their carers, which may be attributed to an education programme.

As the model suggests, learning is conceptualised both as a response to positive reactions to training and as a causal determinant of changes in the trainee's behaviour. Arguably, this linear approach underlies the assumptions that many trainers appear to make about the evaluation of their own teaching. In other words, they collect feedback data from students (learners' reactions), assume that that if the students give positive feedback that they have learned something and they then look out for evidence of good practice by the students in placements, which is in turn attributed to the training[1]. The inadequacies of these assumptions are, I think, self-evident. Nevertheless, the advantage of Kirkpatrick's model is exactly that it does focus attention on possible different levels of evaluation and implies that a comprehensive approach should be concerned with all these levels. Thus, it is insufficient to evaluate training according to whether or not the students enjoyed the presentations and found them informative (the "happy faces" questionnaire), or to assume that it is adequate to establish that students acquired particular skills, in communication, for example, without investigating whether or not they were able to transfer those skills to practice. Further, since the purpose of the whole exercise is to benefit service users and/or carers, a comprehensive evaluation should ask whether training has made any difference to their lives. As I will describe later, the outcomes for users and carers of interventions employed by trainees can be assessed by researchers using standardised measures of, for example, mental health and impairment. Such measures can include users' own ratings of their quality of life as well change in health or problem status. But first we should ask what users themselves consider to be the important outcomes of training.

Thus, it is insufficient to evaluate training according to whether or not the students enjoyed the presentations and found them informative (the "happy faces" questionnaire), or to assume that it is adequate to establish that students acquired particular skills, in communication, for example, without investigating whether or not they were able to transfer those skills to practice. Further, since the purpose of the whole exercise is to benefit service users and/or carers, a comprehensive evaluation should ask whether training has made any difference to their lives.

Users' and carers' views on the outcomes of training

Interestingly, when asked about desirable outcomes of professional education, service users and carers appear to stress Kirkpatrick/Barr Level 2 outcomes regarding attitudes, knowledge and skills rather than Level 4 "benefits" for themselves. For example, user and carer focus groups reflecting on desirable outcomes for the new social work degree emphasised personal qualities such as warmth, empathy and understanding, practical skills, information and

[1]Crisp et al. (2003, p.36) cite a frank example of this approach to assessing the outcomes of a course which confesses that "The authors have relied on their 'gut instincts' as teachers and the ad hoc reports of students and faculty."

the ability to work creatively to find solutions (GSCC, 2002). Similar results were found in the development of a set of user-determined outcomes for the evaluation of a postqualifying course in community mental health, using focus groups and a postal survey of 29 user groups (Barnes et al., 2000). For example, 93% of respondents thought it "very important" that students should treat service users with respect, not as 'labels' and 82% strongly agreed with the statement that, "First and foremost, professionals should develop their capacity to 'be human'". Over three-quarters considered it "very important" that students learned how to involve service users in assessing their needs and 89% agreed that students should "develop knowledge and learn new skills, but should not adopt a 'text book' approach". This last statement seems to imply the need to develop higher level skills such as that of being able "to work creatively" which was mentioned in the GSCC paper.

Specifying and measuring learning outcomes

An important paper by Kraiger et al. (1993) attempted to develop a theoretically based general model of learning outcomes. In effect, what they did was to elaborate significantly Kirkpatrick's Level 2, distinguishing cognitive, skill-based and affective outcomes. Under each of these three headings they classified a number of key variables and suggested how they could be measured. One advantage of this approach is that they can move beyond the definition of basic skills to higher level abilities of the kind we would hope to see as the outcomes of professional education. I shall now apply Kraiger and colleagues' model to social work education and indicate, with reference to empirical studies in social work and health education, how these outcomes may be measured (Table 2).

(1) Cognitive skills

Kraiger et al. (1993) proposed that cognitive skills be classified as verbal (declarative) knowledge, knowledge organisation and cognitive strategies. Thus a student on an interviewing skills course with declarative knowledge should be able to define a concept such as "active listening". This is the sort of outcome traditionally and easily measured in

…user and carer focus groups reflecting on desirable outcomes for the new social work degree emphasised personal qualities such as warmth, empathy and understanding, practical skills, information and the ability to work creatively to find solutions (GSCC, 2002)

written or multiple choice tests. It has been used in training evaluation by, for example, Willets and Leff (2003) who tested psychiatric nurses' knowledge of schizophrenia at the beginning and end of a training course.

The next level would be the development of 'procedural' knowledge and its organisation into a mental map of the process of interviewing comprising a range of key concepts; the more developed the knowledge, the more complex (inter-related) the mental map. We might describe this as the 'internalisation' of knowledge. This kind of knowledge is usually assessed by academic essays, although this procedure is probably not very reliable, even with blind double marking. Its use as an outcome measure is unlikely to be popular with students: imagine asking them to write an essay at the beginning of a module and again at the end!

A promising approach to the assessment of procedural knowledge, which has been explored in medical education, is 'concept mapping' in which students are asked to link a series of concepts in relation to a particular topic. Students are first trained in the concept mapping method and then, before the teaching and without the help of books or papers, are asked individually to draw a map of their existing knowledge. These can then be scored in terms of the structural and relational qualities of the map. Thus West and colleagues (2002) demonstrated that, following training, doctors were able to produce much more elaborate and accurate concept maps about the diagnosis and management of seizures in children than before training. A similar approach could be taken to the measurement of procedural knowledge acquisition in complex professional tasks in social work, such as assessment and initial interventions in child protection.

Concept mapping is a promising approach to the assessment of procedural knowledge.

Table 2: Knowledge, Skills, Attitudes and Behaviour: measuring learning outcomes

Dimension		Measurement
Cognitive	Declarative (verbal knowledge) Procedural (knowledge organisation) Strategic (planning, task judgement)	MCQs; short Concept mapping; case study Probed protocol analysis (interview or interactive DVD)
Skills	Initial skill Compilation of skills Advanced skills (Automaticity)	(Self-ratings); observer ratings (scales) Observer ratings of DVDs of communication skills. Observation (e.g. of assessment interviews)
Affective	Attitudes to users; values Motivational outcomes, self-efficacy	Attitude scales Self-ratings; confidence ratings
Behaviour	Implementation of learning (and barriers)	Self-report; practice teacher/manager report; rating scales
Impact	Outcomes for users and carers	User-defined scales; self-esteem & empowerment; measures of social functioning, mental health, quality of life, child behaviour etc.

Another approach to the measurement of procedural knowledge is to use a manual and employ trained raters to make judgements of students' responses to a case study. Thus Milne et al. (2003) invited trainees to provide open-ended replies to seven standard questions, to be answered in relation to a current client selected from their caseload. Each such reply was scored out of three with reference to the rating manual, giving a score range of 0-21. Higher scores indicate a better knowledge base (in this study, about the formulation in psychosocial terms of the problems of a service user with severe mental illness). The case study method is reported to have good test-related reliability. This approach could be developed to measure changes in students' abilities to conceptualise clients' problems and strengths in other contexts.

Once knowledge has been internalised, we are able to think strategically about its use, a process known as 'metacognition'. Metacognitive skills include planning, monitoring and revising behaviour. An example of high level skills would be reflecting on the process of an interview with a family group so as to modify the worker's alliances with different family members and also think about the overall direction of the interview, while at the same time engaging (cognitively) in active listening with the person who happens to be talking.

Other metacognitive skills include understanding the relationship between the demands of a task and one's capability to perform it. Thus psychological research (cited by Kraiger et al. 1993) shows that experts are generally more able to judge the difficulty of a task than novices, and more likely to discontinue a problem-solving strategy that would ultimately prove to be unsuccessful. These processes may be termed self-regulation and are of obvious importance to the helping professions, including social work.

In social work education, practice assessors are required to make judgements about social work students' metacognitive skills, but it is difficult to know how reliable and comprehensive these assessments might be. The training literature suggests methods such as 'probed protocol analysis' in order to assess trainees' understanding of the necessary steps to solve a problem. For example, electricians would be asked a series of probe questions to investigate how they investigated an electrical fault, e.g. "Why would you run this test, and what would it mean if it fails?", "How would that test help you solve the problem?". Responses to these questions would indicate whether the trainee was generating hypotheses, evaluating evidence, revising plans and so on. There is some evidence of the value of this approach. Thus, Kraiger et al. (1993) reported that experts' ratings of responses to a prior paper and pencil test of students' metacognitive strategies in using the statistical software SPSS were good predictors of exam scores three months later.

In social work education, practice assessors are required to make judgements about social work students' metacognitive skills, but it is difficult to know how reliable and comprehensive these assessments might be.

Probed protocol analysis might have potential as a rigorous approach to measuring social work students' problem solving and critical thinking skills (Gambrill, 1997). One approach might be to train expert raters to ask students probing questions about how they would tackle a constructed case study and score responses using a manual. This would be a development of Milne et al.'s (2002) case study method described above. This method would be expensive to administer, although it could possibly be used for formal and summative course assessments, instead of a traditional essay or exam.

A recent paper by Ford and colleagues (2004) has helpfully elaborated what may be meant by 'criticality'. These researchers describe a case study approach into how learning takes place and they have suggested on the basis of observations of seminars and tutorials that there is some evidence of "progress" to higher levels (p.194). Because the approach is conceptually well grounded, it might well be possible to develop a reliable manualised approach to the assessment of outcomes. Once again this would be quite expensive to use.

Another possibility would be to work with a group of expert practitioners to develop a consensus on the steps necessary to investigate and solve a number of simulated problems and the rationale for these steps. The case simulations could be presented on interactive DVD, allowing possible different approaches to solving the problem. Students could be asked to choose between different steps and the rationales for these. This method would be quite expensive to develop, but inexpensive to operate because scores could be generated automatically. Students could also be given instant (electronic) feedback on their performance which might enhance motivation.

'Probed protocol analysis' may be used to assess trainees' understanding of the necessary steps to solve a problem.

(2) Skills

Skill-based learning outcomes are similarly organised hierarchically by Kraiger and his colleagues (1993). They posit three levels: initial skill acquisition; skill compilation, or the grouping of skills into fluid behaviour; and, through practice,

Kraiger et al. (1993). posit three levels of skill acquisition: initial; skill compilation, or the grouping of skills into fluid behaviour; and, through practice, 'automaticity'.

'automaticity'. Automaticity enables you to accomplish a task without having to think about it consciously and to complete another task at the same time. A familiar example would be the process of learning to drive a car; at the third level you are able to talk to passengers while monitoring road conditions, change gears and react to sudden hazards. We would expect expert social work practitioners to be able to perform certain tasks at a similar level of automaticity. The American social worker Harry Apponte once compared learning the skills of family therapy to learning to box. He suggested that you would know when you had become expert when you "just did it" without having consciously to think about what you were doing. You could then be completely attuned and responsive to what was taking place in the therapy session. In a parallel professional field, Benner (1984) has argued that the expert nurse has an 'intuitive grasp of the situation and zeroes in on the accurate region of the problem without wasteful consideration of a large range of unfruitful alternative diagnoses and solutions ' (p.31-2).

Nerdrum (1997) provides an example of the measurement of initial skill acquisition. Student social workers were invited to suggest helpful answers to ten videotaped statements from simulated clients. The students' written responses were then rated by researchers using a five-point scale of 'empathic understanding' .

A number of studies have asked trainees to rate their own skills before and after training; for example, Bowles et al. (2001) devised a self-report scale to measure communication skills used in brief solution-focused therapy. However the problems with this approach are first that these measures are generally ad hoc and not standardised so we cannot be sure that they measure with reliability and validity. Second, at the beginning of a course trainees may not know how much or how little they know, so later judgements of skills may be compromised. Third, independent observers may not agree with the students' ratings of their skills. (Not all people who think they are good car drivers are considered as such by their passengers.)

A number of studies have asked trainees to rate their own skills before and after training, but ratings by observers offers a more valid & probably more reliable method of measuring initial and compilation skills.

Rating of students' communication skills by observers offers

a more valid and probably more reliable method of measuring initial and compilation skills. For example Cheung (1997) in Hong Kong, had both trainers and trainees assess the content of videotapes of simulated interviews. His purpose was to identify helpful interviewing techniques for use in an interview protocol for social workers and police in child sex abuse investigations. Sanci et al (2000) used both self-ratings of skills and observer ratings on a standardised scale to measure the outcomes of a training programme in adolescent health care for GPs in Australia. The GPs carried out interviews with "standardised patients" – drama students who had been trained to simulate an adolescent with health problems – and also to make ratings of their rapport and satisfaction with the GP interviewers. Generally speaking, the ratings on the different measures were consistent.

Freeman and Morris (1999) in the USA measured higher level compilation skills used by child protection workers in simulated interviews. They employed a coding system to assess the support and information provided by the trainee, as well as the more basic questioning skills. The measure uses samples of interactions between interviewer and interviewee, although in this case, only the interviewer's behaviour was rated. Independent raters were reported to have achieved a very high level of agreement (90%) using the system. Interestingly, in this study although there were improvements in results on a knowledge questionnaire, there was little evidence of improvement in trainees' skills. Freeman and Morris suggested that this difference may be a consequence of the artificiality of the simulated interviews as well as deficiencies in the training programme.

Not surprisingly, the measurement of the highest level of skill development, automaticity, poses significant problems, even when attempting to assess apparently straightforward tasks such as computer programming. Possibly the best indication of automaticity in social work is when students appear, to a trained observer, to have stopped monitoring their own behaviour in the accomplishment of a high level task, or report less conscious awareness of their own actions.

Approaches to the measurement of automaticity in technical skills training use devices such as asking trainees simultaneously to perform a secondary task and/or introducing a distraction when the trainee is (automatically) performing the primary task. Although practitioners might consider that distractions are part and parcel of working life, it is difficult to see how such strategies could be employed in measuring automaticity in professional behaviour. As suggested above, one indicator would be when a trained observer notes that a student is no longer consciously monitoring his or her behaviour while performing a complex task; but this would be difficult to measure reliably.

Benner (1996) has described a procedure for the construction of narrative accounts of nurses' expertise which has been influential also in research on social work (e.g. Fook et al., 2000). Her procedure is summarised, and critiqued, by Nelson and McGillion (2004). Data are collected from nurses' accounts of practice delivered in small peer groups which are facilitated by researchers trained to probe the participants' understandings so as to elicit dimensions of expertise as defined by the model. An important part of the procedure is the careful preparation of participants to engage in the group presentations and it does seem to be successful in enabling professionals to articulate components of their practice which might otherwise remain hidden not only because they have become automatic, but also because they are 'unformalised' (Osmond and O'Connor, 2004). There is much to commend in this approach however there is a risk of imposing a framework on participants. Nelson and McGillion (2004) put this more strongly, arguing that, "Nurses were coached and drilled on the acceptable expertise narrative. Reinforced normative responses were performed by nurses, who articulated expertise, via explicit instructions, and carefully managed group processes." (p. 635). These critics conclude that, "The validity and appropriateness of judging expertise based on first person accounts must be questioned." (p. 637). Certainly, there would be a sound argument for seeking corroborative evidence if this approach were to be employed in outcome research; that would of course be in the best traditions of methodological triangulation.

Benner's (1996) procedure for the construction of narrative accounts of expertise…does seem to be successful in enabling professionals to articulate components of their practice which might otherwise remain hidden not only because they have become automatic, but also because they are 'unformalised' (Osmond and O'Connor, 2004)

(3) Affective (attitudinal) outcomes

The third category of learning outcomes identified by Kraiger et al. (1993) is affectively-based outcomes, including attitudes (Level 2a in Barr et al.'s expansion of the Kirkpatrick framework); this category also includes values and commitment to organisational goals.

Attitudinal outcomes are conventionally measured by means of standardised self-rating scales. For example, Barnes et al., (2000) used the Attitudes to Community Care scale (Haddow and Milne, 1996) to measure and compare the attitudes of a multiprofessional group of students on a postqualifying programme in community mental health. This scale aims to measure attitudes such as 'user-centredness' and commitment to organisational models of community care. Similar Lickert-type scales have been used to measure changes in interprofessional stereotypes and hetero-stereotypes between social work and medical students before and after an interprofessional education programme (Carpenter and Hewstone, 1996).

Kraiger and colleagues also propose 'motivational outcomes', an example of which might be a greater determination to change one's own behaviour in response to learning about racism or about involving cognitively disabled users in planning their own care. Related to this is the idea of 'self-efficacy', that is the (realistic) feeling of confidence that you have the ability to carry out a particular task. This is particularly important in relation to difficult and/or complicated tasks, such as carrying out a complex child care assessment. Good training practice is evidently to break down tasks into component tasks so that trainees can develop competence and confidence before moving on to complex tasks. However, as studies of the implementation of psychosocial interventions have shown, there is a crucial difference between learning a skill in the classroom and using it in practice. For example, Fadden (1997) found that very few of the trainees who completed a training programme in cognitive-behavioural family therapy for schizophrenia actually put their learning into practice with many families. There are always a number of organisational explanations for this common problem.

Kraiger and colleagues concluded that self-efficacy judgements at the end of training were better predictors of scores on subsequent tests than traditional tests of learning.

According to Kreiger et al.'s (1993) review, however, there is good evidence that perceptions of self-efficacy are an important predictor of the transfer of learning to the work setting. Indeed, Kreiger and colleagues concluded on the basis of their own studies that self-efficacy judgements at the end of training were better predictors of scores on subsequent performance tests than traditional tests of learning.

An interesting study of attitudes to service users and of self-efficacy has been reported by Payne et al. (2002). They measured the self-confidence of nurses working for NHS Direct in their own ability to meet the needs of callers with mental health problems. They asked the nurses to consider a number of written case scenarios and rate their confidence to respond adequately using a visual analogue scale. Parallel ratings were made on the Depression Attitude Scale (Botega, 1992) regarding such matters as whether the nurses considered depression to be an illness and whether such patients are 'troublesome'. A very similar approach could be used to assess social work students' confidence in responding to users with mental health problems and other needs, for example, older people.

Sargeant (2000) employed a rather different approach to the measurement of self-efficacy, asking NVQ students whether they believed that they "satisfied the criterion": 'all of the time', some of the time' or 'not at all'. The criteria included generic abilities such as "can deal with unexpected situations" and specific 'care abilities' such as 'responding when clients disclose abuse'. Ideally, trainees should be asked about the extent to which they consider themselves capable of accomplishing a particular task and also their confidence in so doing. These self-ratings could be combined with ratings made by assessors (practice teachers).

(4) Changes in behaviour

Ideally, trainees should be asked about the extent to which they consider themselves capable of accomplishing a particular task and also their confidence in so doing. These self-ratings could be combined with ratings made by assessors (practice teachers).

How can we know whether learning has been implemented? Most studies have relied on follow-up surveys using postal questionnaires or interviews, and in some cases both. For example, in a post course survey of postqualifying award social work students Mitchell (2001) found that former students, and their managers, believed that there had been a

positive effect on the students' practice. Unfortunately the findings from such studies are generally imprecisely reported and may be open to wishful thinking. Stalker and Campbell (1998), in addition to follow-up interviews with postqualifying students on a course in person-centred planning, examined students' portfolios. These suggested that the students had changed in their attitudes and understanding but did not indicate the extent to which they had actually used the methods in practice, i.e. how many service users had been enabled to develop their care plans. We really need harder evidence; potentially more reliable measures involve information on the number of times specific taught interventions have been carried out. The importance of this information is indicated by the generally disappointing findings from implementation studies in mental health services. For example, when Fadden (1997) followed up 59 mental health professionals who responded to a questionnaire about their use of behavioural family therapy, 70% reported that they had been able to use the method in their work. However, the average number of families seen was only 1.7 and a large proportion of these (40%) were seen by a small number of respondents (8%). Of course, asking trainees alone to report on the extent to which they have practised an intervention introduces a potential source of bias because they may not want to let down the trainer; corroboration by a supervisor or manager would be desirable.

The evaluation of changes in behaviour is most straightforward when there is clear evidence as to whether the trainee carried out the learned behaviour or not. For example, Bailey (2002) used a before and after design to monitor changes in assessment for people with interrelated mental health and substance misuse needs. At the start of the course, trainees were asked to complete a proforma on the care they were providing to service users with whom they were currently working. They were subsequently asked to complete a similar proforma for the same clients a month after the training. Of interest was whether or not they had undertaken an assessment in the manner taught on the course; the existence of written assessments could therefore provide clear evidence of the effectiveness of the course in this respect.

How can we know whether learning has been implemented? Most studies have relied on follow-up surveys using postal questionnaires or interviews, and in some cases both.

We really need harder evidence; potentially more reliable measures involve information on the number of times specific taught interventions have been carried out.

Bailey's approach would however be more difficult to apply with social work students. First, unlike practitioners, it may not be possible to measure a baseline if the programme design involves students being taught a skill and then going into practice placements. Second, it might be difficult or impossible to implement the method of working because of the agency or context. For example, it would be possible to collect evidence that students were using task centred casework as taught on the programme (e.g. written, signed contracts setting out service users' goals and tasks, etc). However, a particular student's failure to implement the method may have more to do with the practice agency's function or management than any lack of learning on the part of the student.

Consequently, when evaluating behavioural outcomes, it is important to assess the possible 'barriers' to implementation. One approach here is Corrigan et al.'s (1992) Barriers to Implementation Questionnaire which has been adapted by use in the UK by Carpenter et al. (2003). This measure consists of five subscales, which measure perceived difficulties relating to time and resources, support and interest of managers and colleagues, user and carer beliefs, knowledge, skills and supervision and the trainee's beliefs in psychosocial interventions. Similarly Clarke (2001) concludes a generally pessimistic review of the evidence about the transfer of learning to practice by asserting the importance of determining the factors which are associated with behaviour change following training.

(5) Impact: outcomes for service users and carers

As noted above, when asked to define the desired outcomes of training for social work and social care, service users and carers seem to focus on Level 2 outcomes, changes in attitudes, knowledge and skills (Barnes et al., 2000, GSCC, 2002). Barnes and colleagues (2000) have described the development of a questionnaire to determine user-defined outcomes of postqualifying education in mental health. The questionnaire may be used in confidential postal surveys or structured interviews with an independent researcher. Some findings using this instrument have been presented in Milne

et al. (2003) and Carpenter et al. (2003). In order to assess change, follow up interviews are preferable because the response rate for a repeat survey is low.

From a professional perspective, outcomes for service users and carers are generally considered in terms of changes in such factors as the quality of life, skills and behaviour, self esteem and levels of stress. Standardised instruments have been developed to assess these factors and may be used in assessing the outcomes of training. For example, Milne et al. (2003) have described the use of measures of mental health, life skills and social functioning and psychiatric symptoms of users who were receiving the services of professionals, including social workers, undertaking a postqualifying course in mental health. Leff et al. (2001) assessed changes in carers' 'expressed emotion' (which is associated with relapse in schizophrenia) and hospital readmission rates, comparing the clients of trainees in family therapy with a control group who received education alone. In considering this study, it is worth noting the general point that positive outcomes for clients of education and training in particular interventions should only be expected if the interventions themselves have been shown to be effective (as is the case for family therapy with schizophrenia).

In the field of child care social work, Pithouse et al. (2002) were careful to provide training in evidence-based interventions to foster carers. They used a standardised measure of behavioural problems of the fostered children, rated by the foster carers, and carer self-ratings of stress and responses to the children's 'challenging behaviour'.

From a professional perspective, outcomes for service users and carers are generally considered in terms of changes in such factors as the quality of life, skills and behaviour, self esteem and levels of stress. Standardised instruments have been developed to assess these factors and may be used in assessing the outcomes of training.

…it is worth noting the general point that positive outcomes for clients of education and training in particular interventions should only be expected if the interventions themselves have been shown to be effective…

3. What Do We Want To Know?

At this point, it is perhaps worth clarifying that evaluating the outcomes of social work education can enable us to answer the following questions:

1. Does "it" work? In other words, do students learn the outcomes which we as educators hope they do?
2. Are students able to put their learning into practice?
3. If so, does it make a difference to the lives of service users and carers?

Note that this assumes that we can specify "it", the educational interventions. Second, it assumes that we can check that the intervention is delivered as planned; this is sometimes called "fidelity". One approach to checking fidelity has been developed by Milne and colleagues (2002). This method, called PETS (Process Evaluation of Training and Supervision) involves time-sampling videotape recordings of teaching sessions and the subsequent categorisation by trained observers of the educational interventions used. Thus it is possible to describe the extent to which educators engaged in, for example, didactic presentation versus participatory learning.

Classically, in experimental design the effects of an intervention are assessed in relation to no intervention or the "standard treatment" or usual approach. This implies making a comparison. For example, if we wanted to know whether bringing social work students together with health care students in interprofessional workshops was effective in changing stereotypes we might compare the outcomes with the standard approach of a lecture or presentation on interprofessional working which has been delivered routinely over the previous few years. In practice, therefore, we tend to be interested in a further question:

4. Is Method A more effective than Method B?

In some cases we may explicitly want to test whether one newly designed intervention is actually better than another, "competitor" intervention. The point to make here is that it is best to assume that any intervention is likely to be effective to some degree. Consequently, when comparing interventions we need to be confident that A and B are sufficiently different to have a differential effect.

…evaluating the outcomes of social work education can enable us to answer the following questions:

1. Does "it" work? In other words, do students learn the outcomes which we as educators hope they do?
2. Are students able to put their learning into practice?
3. If so, does it make a difference to the lives of service users and carers?

…when comparing interventions we need to be confident that A and B are sufficiently different to have a differential effect.

4. Research Designs

Having identified outcomes and measures, the next challenge is to develop strong experimental or quasi-experimental designs which are feasible to employ in the evaluation of social work/social care education and training. Potential designs are summarised in Table 3.

Table 3: Possible Research Designs for Assessing Outcomes

Design	Comment
1. Follow up (post test): single group	Useful as formative feedback to the trainers, but cannot inform outcomes.
2. 'Before and after': single group	Quite commonly used, but outcomes cannot be ascribed exclusively to the training intervention.
3. Post-test: two groups	Requires random allocation of students to different conditions.
4. Pre-test, post-test: two groups	Natural comparison groups can be used. Random allocation preferable.
5. Repeated measures, two groups	Students can be randomly assigned to two groups, both of which get the intervention at different times. But requires three measurements.
6. Times series: one group	Requires multiple, unobtrusive observations rather than formal tests.
7. Single-subject experiments	Requires repeated measures of the same person before,during and after the intervention. Small samples.

(1) Post-test only design

The most common form of evaluation in social work education is almost certainly the completion of feedback questionnaires at the end of the course or programme[2]. Such a "post-only" design is useful as formative feedback to the trainers, who can use it to change and develop the course. However, in the absence of information about how much the students knew at the beginning of the course, it cannot tell us about the learning outcomes. The post-only design is therefore inadequate for our purpose and really should be discouraged because it is not that much more difficult to obtain more informative data.

[2]Half the published evaluations of teaching in assessment skills reviewed by Crisp et al. (2003) reported participant feedback only. Similarly, reviews on interprofessional education (Freeth et al, 2002), postqualifying education in mental health (Reeves, 2001) and in-service training in social services (Clarke 2001) found that post-only evaluations predominated.

(2) Pre-test, post-test design

If we are to develop evidence about the impact of social work education, the very least we can do is to assess the students at the beginning as well as at the end of the course. In practice this means asking them to complete the same questionnaire, or take part in a similar interview, on two occasions (conventionally, Time 1 and Time 2). This is not particularly difficult once a suitable measure of outcome has been chosen. For example, Willets and Leff (2003) use a questionnaire to test psychiatric nurses' knowledge of schizophrenia at T1 and T2. Carpenter and Hewstone (1996) measured changes in social work and medical students' interprofessional stereotypes in a similar fashion. This latter study also asked students to rate how interesting and useful they had found the course and, because it had also asked about the students' expectations at T1, the evaluators were able to put the feedback into perspective. For instance, they concluded that some participants' experiences were quite positive given their low expectations; further, because there were two distinct groups of participants, two-factor analysis of variance could be employed to distinguish the differential effects of the programme on the medical and social work students.

Persuading students to complete measures on two occasions does not seem to be difficult if the baseline measure (T1) is presented as part of the introduction to the course. For example, knowledge tests, attitude questionnaires and concept maps (West et al. 2002) are easy to administer in this way. Likewise, students are generally quite willing to complete the measure again at T2 as part of the evaluation of the course, so long as there is time to do this in the final session itself and students hand in their forms before leaving the session. If the measure is straightforward to score, results can easily be analysed and posted on the intranet within a couple of hours. The promise of quick feedback probably enhances motivation to complete the test. Conversely, if no feedback is given, persuading students to complete further questionnaires becomes an increasingly difficult task.

The most common form of evaluation in social work education is almost certainly the completion of feedback questionnaires at the end of the course or programme . Such a "post-only" design is useful as formative feedback to the trainers, who can use it to change and develop the course. However, in the absence of information about how much the students knew at the beginning of the course, it cannot tell us about the learning outcomes.

Persuading students to complete measures on two occasions does not seem to be difficult if the baseline measure (T1) is presented as part of the introduction to the course.

The difficulty with simple pre-post designs, however, is that any changes observed cannot be ascribed exclusively to the

training intervention. For example, improvements may in part be attributed to 'maturational effects', i.e. a general increase in knowledge, skills and improvement in attitudes as a consequence of being engaged in professional training. Second, changes may also be attributed to 'Hawthorne' effects, that is improvements resulting from involvement in an intervention in its own right, irrespective of its content – a simple response to 'something happening'. Third, they may be an artefact of practice in completing the measures; for example thinking more about the answers to questions about knowledge or appreciating the socially desirable responses in attitude scales. Consequently, it is much better to introduce a control group, as will be discussed later. But first we should consider one design which eliminates the effects of maturation or practice.

(3) The post-test, two groups design

In this design, participants are randomly allocated to one of two groups and assessed only at the end of the course. This design may be used to compare two different approaches to learning using any of the outcome measures outlined above. The assumption is that initial randomisation evens out differences, e.g. in initial knowledge, learning ability and motivation, between participants in the two groups. Consequently the difference in the average T2 scores on whatever measure is used gives an indication of which of the two methods is superior. Because the measure is used only once, there is no opportunity for contamination by practice or maturation. Of course, because there is no baseline, what this method is unable to tell us is how much the participants have actually learned. Although this is actually the simplest of all experimental designs, it did not appear in the literature reviewed for this paper. In practice, it would mean persuading a cohort of students to agree to being randomly allocated to one of two methods of learning a topic or skill. This would be easier to achieve if the topic was not formally assessed, which would avoid students' possible anxiety about being at a disadvantage in relation to their colleagues.

(4) Pre-test, post-test: two groups

As noted above, if we are to be reasonably confident that differences between T1 and T2 scores may be attributed to the educational intervention, we need a control group which did not receive the intervention. In terms of the design, this is essentially the same as making a comparison between two different methods of intervention. Ideally, the participants are selected randomly for one of the two interventions, as in the post-test, two group design described above; this is a 'true' experiment. However, 'quasi-experimental' designs are generally easier to accomplish. Here, seemingly equivalent groups which experience different interventions are compared in terms of outcomes. Nerdrum (1997) compared approaches to training in empathy skills by examining outcomes for students on different social work degree programmes in Finland. This approach is potentially very useful for two main reasons. First, it eliminates the problem of some students on a particular programme feeling that they are getting a 'worse' educational intervention than colleagues on the same programme; all get the same intervention. Second, the sample sizes generated for statistical analysis will be much larger if two, or more, whole programmes participate than if one programme's students are divided into two groups. This increases the possibility that the analysis will have sufficient statistical power to detect differences between the two interventions.

Nerdrum (1997) compared…outcomes for students on different social work degree programmes in Finland. This approach is potentially very useful for two main reasons. First, it eliminates the problem of some students on a particular programme feeling that they are getting a 'worse' educational intervention than colleagues on the same programme; all get the same intervention. Second, the sample sizes generated for statistical analysis will be much larger if two, or more, whole programmes participate…This increases the possibility that the analysis will have sufficient statistical power to detect differences between the two interventions.

There are a number of aspects of comparative, between-programme, evaluations which must be taken into account. The first, and most obvious, is that the student groups must be as closely matched as possible in terms of whatever factors are thought to influence learning of knowledge, skills and attitudes. These may include prior education and social work experience, among others. It is probably safest to consider comparison groups as being 'non-equivalent' and to test statistically for socio-demographic differences such as age. There may be differences in the baseline scores on the outcome variables (test scores), i.e. one group performing better on average than the other at T1. In this case, it is possible to adjust statistically for these differences by using analysis of covariance (ANCOVA).

Another crucial factor in cross-programme comparisons is that the time on the programmes at which the interventions are made are similar; it would, for example, be less useful if the comparisons were made between first and second year cohorts. What is more difficult to standardise for is the teacher: some educators are more inspirational or technically proficient than others. In addition to the collection of observational data on what the trainer does, using PETS (Milne et al. 2002) for example, it would also be relevant to ask students to rate the trainer's competence; if there was an apparent difference between the ratings made by the students on different programmes, the potential impact on learning outcomes could be checked statistically using a regression analysis. (Of course, this in itself would be an interesting investigation to carry out.)

On a degree programme it would not be possible to have a 'non-intervention' group. However, there are examples in the literature of comparisons of outcomes between people who received the training and an equivalent group who did not. Thus, Sharples et al., (2003) have reported the evaluation of an 'introduction to management' course which employed a control group of staff from the same local authority who were matched for job responsibility and gender. Unfortunately, from a research perspective, the two groups were found to differ in some important respects because newly appointed managers had apparently been given priority in selection for the course.

Some studies which employ a non-intervention control group are ethically questionable. For example, Palusci and McHugh (1995) in the USA reported the evaluation of a course for paediatricians in the recognition of child sexual abuse. They compared the knowledge of course participants at T1 and T2 with an equivalent group who did not receive the training. This begs the question of whether or not the non-intervention group subsequently received training in such an important part of their work; one rather hopes so. In general, ethical issues in experimental and quasi-experimental study design do not get sufficient attention in published reports, which is not to suggest that these issues were not considered.

On a degree programme it would not be possible to have a 'non-intervention' group. However, there are examples in the literature of comparisons of outcomes between people who received the training and an equivalent group who did not.

It would, for example, have been valuable to read Pithouse and colleagues' views about the assignment of foster carers to treatment and non-intervention groups in order to create matched groups (Pithouse et al., 2002, p. 204).

The literature review undertaken for this paper found no examples of randomised controlled trials of educational interventions in social work education and there are very few examples in health and medical education. Again, difficulties with the assignment of students to groups is probably an important factor.

An interesting example of an approach which could possibly be used in social work education is a study by Cook et al. (1995). They evaluated a two-day programme designed to deliver the basic concepts and techniques involved in delivering community development services to mental health professionals in the US. They used a variation on the pre-test, post-test two groups design to assess the trainees' attitudes towards people with mental illness in the roles of service recipient, service deliverer and trainer. Trainees received the same training on the first day, delivered by someone who was not a user of mental health services. On the second day the 57 trainees were randomly assigned to receive training from either a service user or a trainer who did not have direct experience of using mental health services. There is no mention of the trainees' response to being assigned in this way.

Trainees completed two attitude questionnaires before the first day of training and again at the end of the programme. The authors reported that compared to those who had been trained by a non-user, those who were trained by the user trainer expressed more positive attitudes, towards people with mental illness overall, as service providers and trainers following the training. However, this study, although strong in terms of experimental design illustrates some of the problems of interpretation of findings. The positive change in attitudes reported could be due to the trainees having a different trainer on the second day of the programme. Alternatively, it could be due to some other personal characteristic of the trainer, as opposed to their status simply

The literature review undertaken for this paper found no examples of randomised controlled trials of educational interventions in social work education and there are very few examples in health and medical education.

as a user of mental health services; thus, the generalisations that can be drawn from the study are limited.

Another example of randomisation is demonstrated by a group of medical educators in Germany, although it is rather more difficult to envisage its use in social work education in the UK. Herzig et al., (2003) reported having carried out an earlier study in which they had randomly allocated a group of medical undergraduates to problem based learning (PBL) or conventional lectures as part of a course in pharmacology. These researchers were interested in following up declarative knowledge, assessed by MCQs and short essay questions, within the two groups. The methods used included inviting students in the final year of their course to take a one-hour 'unannounced' exam on which they were promised formative feedback. The other incentive for participation was the offer of a free dinner to the three students who obtained the highest marks, plus another three selected at random! The conclusions which may be drawn from this study are significantly limited by the observation that only 32 of the 112 students who had participated in the original experiment took part in this follow up. Further, as the authors concede, the PBL input, measured in terms of exposure to the method, was actually very little.

A more substantial study, designed to evaluate an educational programme for GPs in adolescent health has been reported by Sanci et al. (2000) in Australia. The GPs were invited, through advertisements, to apply for a free six week university programme of 2.5 hours a week which conferred continuing professional education credits. Applicants were grouped into eight geographical clusters which were randomly allocated to the intervention or control condition, four clusters in each. Note that this randomised cluster design with only eight clusters is not as strong as a randomised participant design because the GPs working in different areas may have differed in important but unknown ways, for example experiences derived from the patient populations of their practices. However, an advantage of this approach is that there was probably less chance of the members of the control group being 'contaminated' by knowledge gained second hand from a colleague in the education group.

Sanci et al.'s (2003) study is interesting in terms of the range of methods used to assess the learning outcomes and the fact that these measures were made at baseline and with two follow ups at 7 months and 13 months. The measures included observer ratings of videotapes of the GPs' interviewing skills. These involved 'standardised adolescents' (sic) - actually drama students coached to present adolescent health problems. These students also completed rating scales to assess the GPs' rapport with them and their satisfaction with other aspects of the simulated consultation. In addition, the GPs completed MCQs and short answer questions (declarative knowledge) and questionnaire measures of perceived self-efficacy.

The authors helpfully provide information about recruitment to the study. Thus, there were 264 expressions of interest in the course (a little over 10% of those who received the mail shot). Of these, just over half (139) agreed to be randomised. On being notified whether they had a place on the programme, 17% of those in the intervention group withdrew, as did 27% of those who had not been allocated a place. The authors do not say whether there were any incentives offered to the members of the non-intervention group, but it was impressive to see that 73% (54 GPs) did agree to complete the questionnaire and the simulated interviews without participating in the programme. In all, 95% of the GPs completed all phases of the evaluation. The findings of this study were very positive, with the education group showing substantially greater improvement in their scores on almost all measures than the control group. Furthermore, the 13 month follow up showed that the improvements had been maintained. However, as the authors acknowledge, improved performance in a simulation does not necessarily translate into better practice in the surgery or benefits for patients.

There are a number of controlled evaluations of the outcomes for service users of training mental health professionals in psychosocial interventions (PSI). The simplest of these designs could potentially be replicated in social work education. For example, Brooker et al. (1992) compared the

outcomes for mental health service users and their carers of interventions delivered by psychiatric nurses on a PSI course with the outcomes for a comparison group of users who were on the caseloads of colleagues in the same service who had not received the training. The researchers used a number of measures which showed that the users in the experimental group improved more in terms of social functioning, and reduced psychiatric symptoms, while the mental health of their carers improved and carers' satisfaction with services increased. It is possible to imagine training social work students in similar psychosocial interventions, for example task-centred practice, solution-focused or cognitive behavioural methods and comparing the outcomes for users and carers with a control group of users from within a practice agency who received the 'standard' service. The weakness of this design, however, is that we could not be sure whether the outcomes were associated more strongly with the personal qualities and enthusiasm of the students or with the range of skills and knowledge they had learned. Nevertheless, it could well be argued that the personal factors were as important an outcome of the training as the technical knowledge and skills. Further, it would be possible to ask users to assess the practitioners' relationships skills using a measure of therapeutic alliance and to compare students and agency workers on this measure.

A similar study by Leff and colleagues (2001) had the same trainees delivering either family therapy or a brief educational intervention to randomised groups of families. This design got over the possible problem of comparing motivated students with possibly jaded agency staff. Those families who received family therapy improved more than those who only had the teaching sessions. It might be argued that this simply shows that family therapy is better than family education, a finding which has previously been demonstrated in a number of studies. However, the fact that these improved outcomes had been achieved indicates that the students must have learned something from the course. Ideally, this would have been established by doing an independent check of the trainees' behaviour, for example by videotaping the family sessions and evaluating the extent to

It is possible to imagine training social work students in similar psychosocial interventions, for example task-centred practice, solution-focused or cognitive behavioural methods and comparing the outcomes for users and carers with a control group of users from within a practice agency who received the 'standard' service.

which they provided the structured interventions taught on the course.

It would be worth seeing if a randomised comparative study design such as that used by Leff et al. (2001) could be used in social work education. It may be noted that the number of trainees in the Leff study was only 17 and that they worked with on average two families each, 30 families altogether. This sample size has sufficient statistical power to detect significant differences between the experimental and control groups on the chosen measures.

A replication of Leff and colleagues' design would require that social work students gained sufficient skills in an intervention which had been shown in previous studies to be effective. The outcomes following training would then be compared with an intervention which was known to be at least benign. We might for example compare the outcomes for carers of participation in 'parent training' groups run by trained social work students with a low key 'drop in' session facilitated by the same students. This could however raise ethical objections on the grounds that some service users were being denied a service which was known to be effective; this kind of concern is commonly raised by practitioners even in cases where there is no clear evidence that the experimental intervention actually works. An alternative approach is to use a 'waiting list control'.

Waiting list controls with repeated measures

If it is unreasonable or impossible to deny an intervention or training which may be beneficial to users or students, then a 'waiting list' control may be acceptable. Here, all participants are given a baseline assessment (T1) and then divided at random into two groups (Table 4). Group 1 receives the intervention, for example a three week intensive module in communication and interviewing skills, and all members are reassessed at the end (T2). Group 2, the (waiting list) controls then start their course and are assessed at the beginning (T2) and the end (T3). Group 1 students are also re-assessed at T3. If the training was successful, we would expect a greater improvement in mean scores between T1 and T2 for Group 1

If it is unreasonable or impossible to deny an intervention or training which may be beneficial to users or students, then a 'waiting list' control may be acceptable.

than for Group 2 (we might anticipate some improvement in Group 2 because of practice effects and other generalised learning on the programme.) However, we would expect a greater increase in mean scores in Group 2 between T2 and T3, i.e. while they were receiving the training, than for Group 1 (although once again we might anticipate a further small improvement in this group on account of continued non-specific learning). If these assumptions proved correct, we could reasonably conclude that there was consistent evidence of improvement associated with the training.

Table 4 Example of a 'waiting list' controlled design for the evaluation of a training intervention.

	T1 (Baseline)	3 weeks	T2	3 weeks	T3
Group 1	Assessment	Training intervention	Assessment	Other studies	Assessment
Group 2	Assessment	Other studies	Assessment	Training intervention	Assessment

Note that the number of repeated measures need not stop at three. It would be desirable to see whether the effects of the training persisted and whether, for example, communication skills were given a boost during the practice placements on a programme, i.e. T4 and T5 measures. Thus T4 measures might be used as part of the 'fitness to practice' assessment required by the GSCC before students go on placement and T5 measures be taken on the completion of that placement. So long as the measures themselves were sufficiently engaging for the students, such that they felt that there were learning from the monitoring of their performance and that the assessments were not too onerous, they might well persist. In contrast, attempts to persuade students to complete the same questionnaire for a fifth time are much less likely to be successful, especially if the measure is perceived to be of marginal relevance to their learning.

So long as the measures themselves were sufficiently engaging for the students, such that they felt that there were learning from the monitoring of their performance and that the assessments were not too onerous, they might well persist. In contrast, attempts to persuade students to complete the same questionnaire for a fifth time are much less likely to be successful, especially if the measure is perceived to be of marginal relevance to their learning.

A more parsimonious approach in terms of measures would employ 'counter balancing'. Here the participants would be randomised into two groups. Group 1 would be trained in

Method A, e.g. brief solution-focused interventions while at the same time Group 2 would learn Method B, e.g. groupwork. Then, the groups would cross over, with Group 2 learning Method A and Group 1 learning Method B. Both groups would be assessed at the end of each module. This approach aims, through randomisation, to deal with 'order effects', associated with practice and generalisation of learning on the one hand and fatigue on the other. Despite the advantages of this design, it does not appear in the literature, possibly because it appears complex and statistical analysis demands slightly more sophisticated analysis of variance techniques. However, training small groups of students in intervention methods seems to be quite common on social work programmes and there may be potential for the use of this design.

(5) Time series designs

Time series designs do not require a control group. Instead they need multiple measures of the group members who receive the educational intervention, including multiple baseline measures. Conclusions about the effects of the intervention are based on an analysis of trends before, during and after the intervention. In more sophisticated designs, the intervention is withdrawn and subsequently reintroduced and the effects noted. Apocryphal stories allege that lecturers have been successfully trained by their students to stand in a particular position in the lecture theatre or to engage in a certain mannerism. The intervention in this case might be smiles and nods of interest which are withdrawn when the desired behaviour disappears. Not surprisingly, these designs are generally associated with operant conditioning (pigeons pecking at discs in "Skinner" boxes). They are often employed in the assessment of interventions for users with severe learning disabilities, but there is no reason why they should not also be used in the evaluation of students trained to use these methods. The focus in this case would be on the students' behaviour (the smiles and nods) rather than that of the service user. Note that in this approach, the numbers of participants need only be few. The argument about the effect of a training intervention relies on the repeated demonstration of the same pattern in a small number of

Time series designs do not require a control group.

individual case studies rather than the aggregated measure of a group. Once again, there do not appear to be any examples of these designs in the literature.

5. Some Practical Considerations

In this section of the paper I suggest some practical considerations, including thoughts about how to engage students in the evaluation of their learning, how evaluation might be linked to assessment and the potential for collaboration between programmes.

(1) Engaging students

I start from the assumption that it is both desirable and essential to engage students in the systematic evaluation of their own learning. In most, if not all, universities and colleges students are required to complete evaluation forms at the end of course modules. These are generally 'smiley faces' measures of interest and enjoyment, satisfaction with the venue and the teaching, perhaps with opportunities to comment on the adequacy or otherwise of the book list and library. At my own university, staff are obliged to ensure a minimum 66% response rate and to report this to the Dean, along with a summary of the results. As I have indicated above, this one group, post-test only design may be useful as formative feedback, but it does not tell us anything useful about outcomes. Further, the findings may not even be reported to the students who completed the forms and, if they are, this may not be done in a useful way. Such practices are, I suspect, all too common. They do of course run entirely against the good practice which is no doubt taught in the research methods classes on the same programme (which are themselves generally evaluated in the same deficient manner, I fear).

Consequently, the first step must be to involve students with staff in any group which is established to assess outcomes systematically. The students must be engaged in discussions about the desirability and feasibility of the various approaches which might be considered. They will certainly want to consider the implications for their time and their learning and to be confident that the findings will be used, and used appropriately. Thus, they might be engaged if they considered that the findings would be reported carefully to them as individuals as well as to the group and that they could use the information to monitor their own performance

I start from the assumption that it is both desirable and essential to engage students in the systematic evaluation of their own learning. Consequently, the first step must be to involve students with staff in any group which is established to assess outcomes systematically. The students must be engaged in discussions about the desirability and feasibility of the various approaches which might be considered.

and learn how to improve. Similarly, students could be interested in the systematic collection of self-report data for use in the portfolios which are required on many programmes. These could include ratings of self-efficacy in a range of relevant practice skills, remembering that self-efficacy ratings are a good predictor of behaviour.

Test results, e.g. MCQs, concept mapping scores and observer ratings of communication skills DVDs could all be used in this way. In some cases students could score these test themselves, or if they completed a test e.g. a case study simulation or knowledge test on a computer the programme could generate an automatic individual score while feeding an anonymous score to a database for group analysis. If the results are to be used in summative as well as formative assessment, the various measures could be included in the range of assessment methods which feed into the overall score for a 'big fat' course module. Such an approach would satisfy the appropriate requests of external examiners and validating panels that professional courses in particular should do more than assess by essay.

(2) Involving service users and carers

Many of the same arguments may be applied to the involvement of service users and carers in the design of systematic evaluations. I have already discussed in general terms the kinds of outcomes in which users and carers express interest. This discussion should take place with users and carer consultants to individual programmes; this may come about through the planned review of a course module and/or practice placements. Users and carers would also be able to advise on the appropriateness and feasibility of proposed measures and research designs. In addition, their contribution to the interpretation of findings, particularly from practice, would be invaluable.

If the results are to be used in summative as well as formative assessment, the various measures could be included in the range of assessment methods which feed into the overall score for a 'big fat' course module.

Users and carers would also be able to advise on the appropriateness and feasibility of proposed measures and research designs. In addition, their contribution to the interpretation of findings, particularly from practice, would be invaluable.

(3) Engaging staff

My impression is that on university-based programmes, much evaluation already happens; the problem is that it is too limited in scope and design so that the information it provides is of very little use. My position is that evaluation

should as far as possible be built into the design and re-design of modules and programmes and seen as part of students' learning. Thus, I am suggesting that the systematic collection of test data becomes part of assessment as well as providing evidence for the evaluation of the module for the university (to this extent they should replace the smiley faces). The assessment of learning outcomes can provide not only case studies for research methods training but also, fundamentally, an enthusiasm for the collection of data and its analysis within the general framework of evidence-based or, as I prefer, research-informed practice.

To put this another way, the collection and analysis of course evaluation data could be that much more interesting, and much less of a chore, if programme staff (1) used a wider range of measures which could provide information about the learning outcomes in which they were interested, and (2) attempted some of the research designs suggested in this paper, so that they could have evidence to see whether the learning outcomes had been achieved.

As a first step it would be well worth examining the learning outcomes for course modules and reviewing how these might be measured. Since learning outcomes are often poorly specified, this exercise would be valuable in itself: if they are to be measured they will have to be revised in order to be observable. Similarly, this process would help improve the alignment of learning and teaching methods with the learning outcomes (Biggs, 1999). The next step would be to assess these at the beginning and end of the module. The results of this pre-test, post test single group design has limitations, as noted above, but it would at least give an indication as to whether the outcomes were being achieved. The third step would be to use some of the stronger designs discussed above.

The incremental approach which I am proposing could at a later stage attempt to put all these pieces together to examine the extent to which the design of the curriculum as a whole enabled the achievement of 'holistic' learning outcomes which in effect transcend individual modules (Burgess, 2004).

…evaluation should as far as possible be built into the design and re-design of modules and programmes and seen as part of students' learning.

(4) Collaboration between programmes

As I have suggested above, there would be significant advantages in programmes collaborating in evaluation studies. This would both allow the comparison of different methods of teaching and learning and the creation of experimental and control/comparison groups which would be large enough to generate sufficient statistical power to detect significant differences between conditions. But there is a further reason for collaboration: this is because we know so little about how these methods of evaluation might be put into practice. I believe that it would only be through enthusiasts sharing ideas and experiences that we could hope to make progress.

There would be significant advantages in programmes collaborating in evaluation studies. This would both allow the comparison of different methods of teaching and learning and the creation of experimental and control/comparison groups.

6. Conclusion

SCIE (2004) has called for increased funding for the evaluation of social work education after (yet) another knowledge review "…revealed that social work educators faced a lack of evidence in deciding which teaching and learning methods are effective." In this discussion paper I have attempted to outline how this evidence might be accumulated: I hope to have identified a range of relevant outcomes of social work education and indicated some appropriate approaches to their measurement. I have also reviewed a number of research designs which might prove feasible for use on social work programmes at both qualifying and postqualifying levels.

One way forward would be a facilitated learning set approach involving a group of programmes, plus user and carer consultants. The group would review and agree to try out different methods and designs, share their experiences, review and refine promising approaches. Following this, participants would set up comparative studies within and between programmes and share in the analysis and dissemination of findings. In this way, we might reasonably hope both to build capacity and capability amongst academics and trainers (including users and carers) in the evaluation of social work education and to generate high quality evidence about the effectiveness of methods of teaching and learning. We surely owe this to our students, our colleagues and most importantly, the people who will receive the services which our students will provide.

One way forward would be a facilitated learning set approach involving a group of programmes, plus user and carer consultants. The group would review and agree to try out different methods and designs, share their experiences, review and refine promising approaches.

7. References

Bailey, D. (2002b) Training together—part two: the evaluation of a shared learning programme on dual diagnosis for specialist drugs workers and Approved Social Workers (ASWs). Social Work Education, 21, 685-699.

Bailey, D., Carpenter, J., Dickinson, C. and Rogers, H. (2003) Post Qualifying Mental Health Training. London, National Institute of Mental Health, www.nihme.org.uk

Barnes, D, Carpenter, J, Bailey, D (2000) Partnerships with service users in interprofessional education for community mental health: a case study. Journal of Interprofessional Care, 14, 191-202.

Barr, H., Freeth, D., Hammick, M., Koppel, I. & Reeves, S. (2000). Evaluating Interprofessional Education: a United Kingdom review for health and social care. BERA/ CAIPE.

Benner, P. (1984) From Novice to Expert. Excellence and Power in Nursing Practice. Addison-Wesley, Menlo Park, CA.

Benner, P. (1986) Appendix A in Expertise in Nursing Practice: Caring, Clinical Judgement and Ethics, New York, Springer Press.

Biggs, J. (1999) Teaching for Quality Learning at University, Buckingham, Open University Press.

Botega, N, Mann, A, Blizzard, R and Wilkinson, G. (1992) General Practitioners and depression – first use of the Depression Attitudes Scale. International Journal of Methods in Psychiatric Research, 2, 169-180.

Bowles, N., Mackintosh, C. and Torn, A. (2001) Nurses' communication skills: an evaluation of the impact of solution focused communication theory. Journal of Advanced Nursing, 36, 347-354.

Brooker, C., Tarrier, N., Barrowclough, C., Butterworth, A. & Goldberg, D. (1992). Training community psychiatric nurses for psychosocial intervention: report of a pilot study. British Journal of Psychiatry, 160, 836-844.

Burgess, H. (2004) Redesigning the curriculum for Social Work Education: complexity, conformity, chaos, creativity, collaboration? Social Work Education 23, 163 – 183

Carpenter, J. and Hewstone, M. (1996) Shared learning for doctors and social workers. Evaluation of a programme. British Journal of Social Work, 26, 239-257.

Carpenter, J., Barnes, D. and Dickinson, C. (2003) Making a Modern Care Force: evaluation of the Birmingham University Programme in Community Mental Health. Durham: Centre for Applied Social Research. www.dur.ac.uk/sass/casr/

Cheung, K.M. (1997) Developing the interview protocol for video-recorded child sexual abuse investigations: a training experience with police officers, social workers and clinical psychologists in Hong Kong. Child Abuse and Neglect, 21, 273-284.

Clarke, N. (2001) The impact of in-service training within social services. British Journal of Social Work, 31, 757-774.

Clarke, N. (2002) Training care managers in risk assessment: outcomes of an in-service training programme. Social Work Education, 21, 461-476.

Corrigan, P.W., Kwartarini, W.Y & Pramana (1992). Barriers to the implementation of behaviour therapy. Behaviour Modification, 16, 132-144.

Crisp, B. R., Anderson, M. R., Orme, J. and Green Lister, P. (2003) Learning and Teaching in Social Work Education: Assessment. Bristol: The Policy Press.

Fadden, G. (1997) Implementation of family interventions in routine clinical practice following staff training programmes: a major cause for concern, Journal of Mental Health, 6, 599-612.

Fook, J., Ryan, M. and Hawkins, L. (2000) Professional Expertise: practice, theory and education for working in uncertainty. London, Whiting and Birch.

Ford, P., Johnstone, B., Mitchell, R. and Myles, F. (2004) Social work education and criticality: some thoughts from research, Social Work Education, 23, 185-198.

Freeman, K. A. and Morris, T. L. (1999) Investigative interviewing with children: evaluation of the effectiveness of a training program for child protective service workers. Child Abuse and Neglect, 23, 701-713.

Freeth, D., Hammick, M., Koppel, I., Reeves, S. and Barr, H. (2002) A Critical Review of Evaluations of Interprofessional Education. London: LTSN for Health Sciences and Practice, King's College.

Gambrill, E. E. (1997) Social Work Practice. A Critical Thinker's Guide. New York, Oxford University Press.

General Social Care Council (2002) Reform of the Social Work Degree: Reform Focus Groups Summary. 4 September 2002. www.gscc.org.uk

Haddow, M & Milne, D. (1995) Attitudes to community care: development of a questionnaire for professionals. Journal of Mental Health. 4, 289-296.

Herzig, S. , Linke, R. M., Marxen, B., Börner, U. and Antepohl, W. (2003) Long-term follow up of factual knowledge after a single, randomised problem-based learning course. BMC Medical Education, 3, <http://www.biomedcentral.com/1472-6920/3/3>.

Hulsman, R. L., Ros, W. J. G., Winnubust, J. A. M. and Bensing, J. M. (1999) Teaching clinically experienced physicians communication skills. A review of evaluation studies. Medical Education, 33: 655-668.

Kirkpatrick, D.L. (1967). Evaluation of Training. In R.L. Craig, & L. R. Bittel, Training and Development Handbook (pp.87-112). New York: McGraw-Hill.

Kraiger, K. Ford, K. and Salas, E. (1993) Application of cognitive, skill-based and affective theories of learning outcomes to new methods of training evaluation. Journal of Applied Psychology, 78, 311-328.

Leff, J., Sharpley, M., Chisholm, D., Bell, R. and Gamble, C. (2001) Training community psychiatric nurses in schizophrenia family work. A study of clinical and economic outcomes for patients and relatives. Journal of Mental Health, 10, 189-197.

Lyons, K and Manion, K. (2004) Goodbye DipSW: trends in student satisfaction and employment outcomes. Some implications for the new social work award. Social Work Education, 23, 133-148.

Marsh, P. and Triseliotis, J. (1996) Ready to Practice? Social Workers and Probation Officers: their training and first year in work. Avebury, Aldershot.

Milne, D.L., James, I., Keegan, D., Dudley, M. (2002) Teachers' PETS: a new observational measure of experiential training interactions. Clinical Psychology & Psychotherapy, 9, 187-199.

Milne, D., Carpenter, J., Lombardo, C. and Dickinson, C. (2003) Training for Evidence Based Practice in Mental Health, York, Northern Centre for Mental Health.

Mitchell, C. (2001) Partnership for continuing professional development: the impact of the Post Qualifying Award for Social Workers (PQSW) on social work practice. Social Work Education, 20, 434-445.

Nelson, S. and McGillion, M. (2004) Expertise or Performance? Questioning the rhetoric of contemporary narrative use in nursing. Journal of Advanced Nursing, 47, 631-638.

Nerdrum, P. (1997) Maintenance of the effect of training in communication skills: a controlled follow-up study of level of communicated empathy. British Journal of Social Work, 27, 705-722.
Osmond, J. and O'Connor, I. (2004) Formalizing the unformalized: practitioners' communication of knowledge in practice, British Journal of Social Work, 34, 677-692.

Palusci, V. J. and McHugh, M. T. (1995) Interdisciplinary training in the evaluation of child sexual abuse. Child Abuse and Neglect, 19, 1031-1038.

Pawson, R., & Tilley, N. (1997) Realistic Evaluation. London: Sage Publications.

Payne, F., Harvey, K., Jessopp, L., Plummer, S., Tylee, A. and Gournay, K. (2002) Knowledge, confidence and attitudes towards mental health of nurses working in NHS Direct and the effects of training. Journal of Advanced Nursing, 40, 549-559.

Pithouse, A., Hill-Tout, J. and Lowe, K. (2003) Training foster carers in challenging behaviour: a case study in disappointment? Child and Family Social Work, 7, 203-214.

Sanci, L. A., Coffey, C. M. M., Veit, F. C. M., Carr-Gregg, M., Patton, G. C., Day, N. and Bowes, G. (2000) Evaluation of the effectiveness of an educational intervention for general practitioners in adolescent health care: randomised control trial. British Medical Journal, 320: 224-230.

Sargeant, A. V. (2000) An exploratory study of the effects of progression towards National Vocational Qualifications on the occupational knowledge and care practice of social care workers. Social Work Education, 19, 640-661.

Sharples, A., Galvin, K., Holloway, I. and Brown, K. (2003) The impact of training: an evaluation of an 'introduction to management' course for social services staff. Learning in Health and Social Care, 2, 37-50.

Social Care Institute of Excellence (2004) SCIE calls for research funding to make social work education evidence-based. Media release. 27th July.
Stalker, K. and Campbell, V. (1996) Person-centred planning: an evaluation of training programme. Health and Social Care in the Community, 6, 130-142.

Trevithick, P., Richards, S., Ruch, G., Moss, B., Lines, L. and Manor, O. (2004) Learning and Teaching Communication Skills on Social Work Qualifying Courses/Training Programmes. Bristol, Policy Press.

West, D. C., Park, J. K., Pomeroy, J. R., Sandoval, J. (2002) Concept mapping assessment in medical education: a comparison of two scoring systems. Medical Education, 36, 820-826.

Willetts, L. and Leff, J. (2003) Improving the knowledge and skills of psychiatric nurses: efficacy of a staff training programme. Journal of Advanced Nursing, 42, 237-243.

Evaluating Outcomes in Social Work Education

The aim of this paper is to stimulate discussion amongst educators and evaluators by attempting:

1. To identify what we mean by the 'outcomes' of social work education
2. To consider how these outcomes might be measured
3. To assess the advantages and disadvantages of different research designs for the evaluation of outcomes in social work education
4. To illustrate some of the methods and measures which have been used to evaluate outcomes

ISBN 0-9549544-0-8

www.sieswe.org
www.scie.org.uk