PART I

PSYCHOTHERAPY AND COUNSELLING OUTCOME RESEARCH

Despite the fact that psychotherapy and counselling are scientific methods of treating psychological and psychiatric problems and disorders, they are also considered as a sort of social care that is hardly recognised as a treatment. This is stressed, though less so in recent years, by some opponents of psychotherapy and counselling outcome research. The fact that what is happening in the therapeutic relationship cannot be captured in naturally reductive scientific inquiry leads to scepticism about the value of outcome and other research. Furthermore, studying therapeutic outcomes is potentially threatening for therapists, if it shows that their work is not that effective. So why bother with studying whether psychotherapy works?

There may be several reasons. In practice, psychotherapists often form their own creative ways of working with clients. Many psychotherapeutic theories were developed on the basis of clinical experience, reflection and cautious reasoning. Counselling and psychotherapy outcome research allows for the assessing of such theories. Similarly, outcome research assesses both approaches that exist and approaches that are developed as a variation on existing therapies due to the limitations of the original approaches with certain types of client. Outcome research may also assess approaches that are developed on the basis of process or other psychological research. Generally speaking, outcome research assesses therapeutic approaches that researchers consider to be potentially effective and therefore worth studying. It is a means of evaluating and validating the effectiveness and efficacy of psychotherapy and counselling. Outcome research often assesses the effectiveness and efficacy of a psychological treatment against alternative forms of treatment, such as pharmacotherapy or self-help groups. Costs are taken into consideration in these instances too.

Psychotherapy and counselling outcome research involves many stakeholders. Founders of different therapeutic approaches are among them. They, as well as the therapists trained in their respective approaches, are substantially interested in having their approach empirically validated so that they can gain security in the therapy market. Other stakeholders are the training providers in the various therapeutic approaches. In many countries, the trend is to fund only the training that provides empirically based treatments. Outcome research is very

relevant to another group of stakeholders – practising therapists – who may want to know what kind of approach could be promising in their work with a particular client.

Other potential stakeholders are politicians or insurance companies that decide what kind of treatment will be provided in state-funded or insurance-funded medical care. These stakeholders will naturally be cautious when assessing the effectiveness and efficacy of psychotherapy and counselling, as the funds for medical care are always tight and come under many competing pressures. Last but not least, an important group of stakeholders are clients, the consumers of psychotherapy and counselling. They are definitely entitled to know what kinds of effects they can expect from any psychological treatments they are about to undergo.

All these different expectations, interests and pressures influence how psychotherapy and counselling outcome research is conducted. There is a demand for high ethical standards so that findings are not consciously or unconsciously distorted.

1

Instruments Used in Psychotherapy and Counselling Outcome Research

What the goal of psychotherapy and counselling should be is often the subject of theoretical debate. For example, some approaches favour improvement in psychopathological symptoms, some changes in interpersonal functioning and biographical self-understanding, and some the pursuit of individuals' potential and personal development. The issue may be further complicated by a particular ethical perspective weighting the impact of different changes achieved in therapy (see Tjeltveit, 1999), e.g. the goals of treatment when working with real guilt.

This chapter will take a pluralistic approach, presenting targets for measuring therapy outcome regardless of their theoretical origin. The main guideline will be 'current' practice and currently used instruments. By this, however, I do not want to underestimate any particular context of how is change understood, which definitely influences how therapy is conducted and studied. I will not focus here on generic issues of measurement such as the reliability or validity of the instruments used (see Kaplan & Sacuzzo, 2005), but rather on issues more specific to therapy outcome.

Measuring therapy outcome is a complex matter not only because of problems with the delineation of areas we want to improve by therapy but also because of the complexity involved in assessing quantitatively whether enough change has occurred. This complexity arises from the fact that different methodological approaches to measuring are differently sensitive to the amount of change.

Furthermore, different therapeutic approaches may be differentially effective in different areas of therapeutic outcome. For example, two therapies for depression may be differentially effective, with one being more effective in reducing symptoms of depression and other more effective in the area of improved interpersonal functioning. Globally, we could say that the outcome measured may not only be the function of therapy, but also the function of the *construct* (*variable*) that is being assessed, its *sensitivity to change* (or the sensitivity of the instrument that was used), the *perspective* taken

(client, therapist, expert, significant other, objective data, etc.), and the *time of assessment* (e.g. at the end of treatment vs. follow-up).

Before we move on to introducing different methods of measuring outcome, we will briefly focus on a quantitative expression of the magnitude of change, the so-called *effect size* (see Cohen, 1988) and criteria that tell us when we can talk about reliable and clinically meaningful change.

Effect Size and the Magnitude of Change

Effect size, in the context of measuring psychotherapy outcome, is a numerical expression of the difference between the means of two or more compared groups as measured by an outcome measure. It includes comparisons of outcomes within the same group, before and after therapy, as well as comparisons of different groups after therapy, e.g. a group of patients receiving psychotherapy and a control group without therapy. Effect size allows us to be more specific about the magnitude of observed change (if it is present) than just simply stating whether the groups differ or not. In psychotherapy research, Glass (see e.g. Smith, Glass & Miller, 1980) adapted Cohen's *d* to allow the magnitude of difference between the experimental and control group to be measured. Mathematically speaking, we are using following formula:

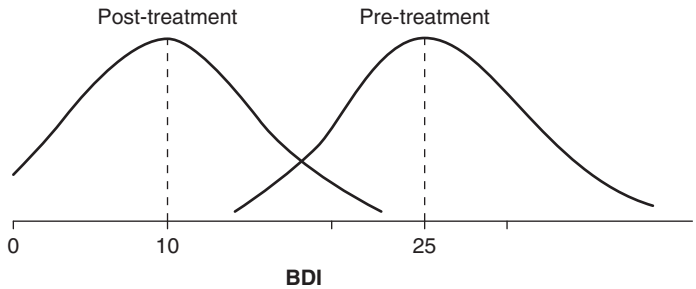
$$ES = \frac{\bar{x}_e - \bar{x}_c}{s}$$

where *ES* means effect size, $\bar{x}_e$ is the mean of the experimental group, $\bar{x}_c$ is the mean of the control group, and the standard deviation *s* is computed as the pooled standard deviation of both groups' distributions (sometimes the standard deviation of the control group is used).

I will illustrate the computation of effect size by using a simple example. Let us suppose that we have a group of depressed patients with the mean score before treatment of 25 on the Beck Depression Inventory (BDI) (the possible range of the score is from 0 to 63) and that the standard deviation of distribution of patients' score is 7. After treatment, the group's score is 10 on the BDI and the standard deviation would, let's suppose, remain the same.

To compare the difference measured by the BDI before and after treatment, we replace $\bar{x}_c$ in the formula above with the group mean before the treatment and $\bar{x}_e$ with the group mean after the treatment. Then we put the pooled mean of both standard deviations into the denominator (to make it easier, let's suppose it would be 7). The effect size is then 25 minus 10 divided by 7, which is 2.14. This is a large difference according to Cohen's categorization of effect sizes in social sciences, where an effect size greater than 0.80 is considered as large, 0.50 as medium and less than 0.20 as small (Cohen, 1988). The example is graphically presented in Figure 2.1.1. As can be seen, the distributions of the score before and after treatment overlap minimally.

Figure 1.1 Distributions of BDI scores of our illustrative example with the pre-treatment mean score 25 (standard deviation 7) and post-treatment mean score 10 with the same standard deviation



The above formula can also be used for comparing two different patient groups, e.g. an experimental and a control group, or two groups in two different active treatments (i.e. two experimental groups). In this case, $\bar{x}_e$ stands for the mean of one group and $\bar{x}_c$ for the mean of another group, with *s* computed as the pooled standard deviation of both groups. Use of the above formula assumes that both groups have almost identical parameters before treatment (means and standard deviations). A more conservative method computes pre-post ESs for both groups separately; the difference between the ESs for the first and second group is then considered the magnitude of difference between groups, hence the ES comparing these groups (see Elliott, 2002b).

Effect size has a broad use. Besides allowing for a more exact estimation of the difference between the groups compared, it also allows the magnitude of effect sizes across several studies to be measured. Thus it can be a basis for the meta-analysis of the cumulative results from several studies investigating outcome (I will address this issue in Chapter 4, which focuses on meta-analysis).

Reliable Change and Clinical Significance

Effect size expresses the magnitude of difference obtained using a particular instrument when comparing two groups, usually experimental and control. It does not, however, tell us whether the outcomes gained by individual patients are also significant in real life. The mean for a particular patient group can, statistically, significantly improve; however, this may not mean that the patients are no longer depressed. Statistically significant change can, for example, be an improvement in the group mean from a score of 25 on the BDI to a score of 20, which would mean that patients are on average still depressed. Therefore, if we want to assess change relevant to patients' real-life situation, we should be asking whether the patients in a particular treatment are not depressed after the treatment.

Whether change occurred, and how big this change is, is usually assessed against two criteria (Jacobson & Truax, 1991: (1) the score of a particular patient on a particular measure must move from the range of dysfunctional population to the **functional** range; and (2) this change must exceed the measurement error. For example, a patient who filled out the BDI repetitively would not always achieve the same scores. The score is related to the reliability of the measure, i.e. the stability of measurement and how this measurement is affected by random errors. To say that an individual patient has really improved, his or her score measured after therapy has to differ from the score before therapy enough to exceed the estimated interval of measurement error. The index determining that the change is sufficiently reliable (i.e. it is not only the effect of random error) is called *Reliable Change Index* (RCI) (Jacobson & Truax, 1991) and is computed according to the following formula:

$$RC = \frac{x_2 - x_1}{S_{diff}}$$

where x_1 means pre-therapy score, x_2 means post-therapy score, and S_{diff} is the standard error of difference between two scores that can be computed using the standard error of measurement according to the formula:

$$S_{diff} = \sqrt{2(S_E)^2}$$

where S_E means standard error of measurement (we can compute it from the test–retest reliability of the measure and standard deviation of normative data). If the reliable change (RC) is greater then 1.96, then the probability that the post-therapy score expresses the real change is high (i.e. it exceeds the conventional interval of 95%).

As I mentioned above, the fact that change is real (not random) is usually not enough. We also need to know if the change is sufficient, i.e. that it is relevant in regard to the problem that was addressed by therapy. We can assess the relevance of change by the fact that the score after therapy not only really differs from the score before therapy, but that it is more likely to belong to the range of the healthy, non-clinical (e.g. not depressed) population than to the clinical (e.g. depressed) one. Jacobson and Truax (1991) suggest three ways of testing that the change which exceeds random error is clinically significant:

(1) The score after therapy should fall outside the range of dysfunctional population, i.e. to the range of two standard deviations beyond the mean for that population (i.e. in direction to the functionality).
(2) The score after therapy should fall within the range of two standard deviations of the mean of the functional non-clinical population.
(3) The score after therapy is closer to the functional population than it is to the dysfunctional population.

Which option we choose depends on what normative data are available and whether the normative data of both functional and dysfunctional populations overlap or not. If the norms of the functional population are not available, we should use option (1), i.e. the patient belongs to the 'healthy' population if the change is reliable (exceeds standard error of measurement) and is two standard deviations beyond the mean of the patients' group or other dysfunctional referential group. If the norms of the functional population are available, then we can use option (2), but if the norms of both functional and dysfunctional groups are available, and distributions of these groups overlap, then we should use option (3). For further details about computations of all three options, see Jacobson and Truax (1991).

Jacobson and Truax's (1991) approach to the calculation of clinically significant change is not the only one that can be seen in the literature. Different methods have been suggested (see Speer, 1992; Speer & Greenbaum, 1995; Tingey et al., 1996; Wise, 2004), leading to slightly different outcomes (see Bauer, Lambert & Nielsen, 2004), which, however, may be more similar given the increasing reliability of the measure (Atkins et al., 2005).

Instruments Used in Measuring Therapy Outcome – Source of Data, Constructs Assessed and Sensitivity to Change

The perspective taken when evaluating therapeutic change has a strong impact on the magnitude of observed change. In other words, whether the outcome is assessed by clients, significant others, therapists, expert raters, etc. influences the size of observed change if it is present. In following pages I focus on different evaluators of therapeutic outcome, the types of assessment they may adopt, and give examples of measures that are used.

Clients' Assessment of Outcome

The easiest and the most feasible way of evaluating therapeutic change is that of the clients' self-reports. These usually employ the format of self-report scales and questionnaires. We can categorize them according to the variables they are attempting to measure. We can distinguish personality questionnaires (measuring personality traits or dimensions), symptom scales, well-being and adaptability scales, instruments focused on the constructs related to mental health or instruments focusing on the quality of interpersonal relationships. There are also idiosyncratic measures, as well as instruments developed specifically for evaluating psychotherapeutic outcome. Recently, qualitative methods (mostly structured interviews; e.g. Elliott, 2002a) examining the impact of therapy have become more popular. Finally, we also encounter questionnaires evaluating clients' satisfaction with counselling or psychotherapy.

Personality Questionnaires

Common personality questionnaires that are routinely used in psychological assessment have also been employed in counselling and psychotherapy outcome research. For example, the *Minnesota Multiphasic Personality Inventory* (MMPI, see e.g. Hill, 1989) or the German *Freiburger Personality Inventory* (FPI, see e.g. Kächele et al., 2001) are often used. Although the MMPI was very often used in outcome research and some of its scales seem to be sensitive to therapeutic change, Lambert and Hill (1994), in their authoritative review of outcome instruments and their sensitivity to change, do not recommend this instrument for measuring therapy outcome (mostly because it is time consuming). Generally, it could be said that the personality questionnaires are not very sensitive to change, because they are focused on relatively stable personality traits (Lambert & Hill, 1994) and so are probably not sensitive to distress that the client is currently experiencing. Where a personality questionnaire is used as an outcome measure, it is probably better to use an instrument that also covers current pathology or distress.

Symptom Self-report Scales

These are the methods that focus on clients' psychopathological symptoms. They are usually very sensitive to change (Lambert & Hill, 1994). The most well-known multisymptomatic tool is the Symptom Checklist-90 (SCL–90) (Derogatis, Lipman & Covi, 1973) and also its revised version, SCL–90–R (see e.g. Derogatis, 2000). SCL–90 is a 90-item self-report scale, also suitable for use as a psychopathological screening tool. It contains nine symptom scales: Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation and Psychoticism, but for psychotherapy research the most often used is the Global Severity Index (GSI), which expresses the overall level of psychopathology. SCL–90 is very often used as an outcome tool and is recommended for heterogeneous patient groups (Lambert & Hill, 1994).

One of the most widely used monosymptomatic instruments is the *Beck Depression Inventory* mentioned earlier (BDI; Beck, Steer & Garbin, 1988). It is a 21-item self-report scale. The items cover cognitive, affective and behavioural symptoms of depression. There exist many other symptomatic scales focused on various psychopathological symptoms (anxiety, obsessive-compulsive, panic, etc.). An overview of many instruments can be found, for example, in Hersen (2004).

Questionnaires Assessing Constructs Related to Mental Health

One example of psychological constructs frequently employed as an indicator of psychotherapy outcome is self-esteem. Self-esteem is very often measured by the *Rosenberg*

Self-Esteem Scale (RSES) (Rosenberg, 1965). Focusing on self-esteem can be appropriate when measuring the outcome of therapy for depression (see, for example Greenberg & Watson, 1998).

Questionnaires Assessing the Quality of Social or Interpersonal Functioning

The quality of interpersonal relationships is strongly linked to the quality of mental health (this is most obvious in psychodynamic theories). One of the most widely used instruments focusing on relationships and social functioning is the *Inventory of Interpersonal Problems* (Horowitz et al., 1988). Broader social functioning is covered for example in the *Social Adjustment Scale* (Weisman & Brothwell, 1976). Questionnaires dealing with social functioning are probably less sensitive to change than instruments that focusing on the psychopathological symptoms of depression or anxiety (see Greenberg & Watson, 1998).

Self-report Scales Assessing the Client's Idiosyncratic Problems

These instruments try to assess whether clients see any change in the issues they found troubling at the beginning of therapy. Generally, these tools are very sensitive to therapeutic change (e.g. Greenberg & Watson, 1998). In fact, they can probably overestimate therapy outcome. For example, problems that bothered the client at the beginning of therapy can become irrelevant due to change in the client's life situation. One of the best-known instruments is probably that of *Target Complaints* (Battle et al., 1966). This instrument can be used as a general basis for an interview about the client's most troublesome problems rather than just as a simple self-report scale. The client and the therapist may work together on the formulation of the client's target complaints and the extent to which they are concern him or her (see a version of the instrument in Box 1.1.).

Another similar tool is Shapiro's *Personal Questionnaire* (Philips, 1986; Elliott, Mack & Shapiro, 1999). This instrument, in comparison to the Target Complaints, allows not only formulation of the client's presenting issues at the start, but also the addition of problems that occur later in the course of psychotherapy. It also allows for an evaluation of how long these problems had lasted before therapy started. Like Target Complaints, this instrument may be used collaboratively with the client. The collaborative interview should then be conducted, in the research context, by a trained person other than the therapist to avoid the client's reactivity. The Personal Questionnaire may be used regularly before each session to evaluate the client's difficulties since the last session (see e.g. Elliott, 2002a).

Box 1.1 An example of a version of the Target Complaints (Battle et al., 1966)

Instruction: Please list three target complaints (problems) that bothered you recently (if there are more, you can include them all). Mark the scale according to how much they distressed (bothered) you.

1 _____

0	1	2	3	4	5	6	7	8	9	10	11	12
not at all						very much						could not be worse

Note: Battle et al. (1966) recommend a vertical line with blank boxes without numbers.

Self-report Questionnaires Developed Specifically as Outcome Measures

For a long time psychotherapy researchers were trying to develop a standard battery of instruments that could be used to adequately and sensitively assess therapeutic change (see e.g. Waskow & Parloff, 1975; Strupp, Horowitz & Lambert, 1997). One product of such endeavours has been the development of questionnaires serving specifically to measure counselling and psychotherapy outcome. These questionnaires measure areas that proved to be sensitive to therapeutic change, such as well-being, psychopathological symptoms and life functioning (Howard, Lueger et al., 1993). One of the best-known instruments developed to measure therapy outcome is the *Outcome Questionnaire-45* (OQ–45) (Lambert, Morton et al., 2004); another is the *Clinical Outcomes in Routine Evaluation – Outcome Measure* (CORE–OM) (Barkham, Margison et al., 2001).

The OQ-45 contains 45 items that cover three specific domains: symptom distress, interpersonal relationships and social role performance. The instrument is currently widely used in so-called *patient-focused research* (e.g. Lambert, Hansen & Finch, 2001), which we will address further in Chapter 6. It comes with software enabling the therapist to benchmark the client's progress from session to session against successful as well as unsuccessful clients who filled out the questionnaire before.

The CORE–OM questionnaire developed by the British authors is becoming popular, as it is quite user friendly (e.g. the use of the questionnaire in hard copy is copyright free, which enables practitioners to easily photocopy the instrument). Thirty-four items of the CORE–OM cover four domains: subjective well-being, problems/symptoms, life functioning, and risk/harm. The first three domains, like the domains in the OQ–45, correspond to the phase model of change (Howard, Lueger et al., 1993) that we will discuss further in Chapter 6. The popularity of this questionnaire is also based on the possibility of benchmarking against a huge national database (Mellor-Clark et al., 2006). The instrument is a part of a broader CORE package that consists of pre-therapy and after-therapy forms of collecting information on demographics, diagnosis, type of therapy provided, etc. An example of the items of the CORE-OM is in Box 1.2.

Box 1.2 An example of instructions and items of the CORE-OM (Barkham et al., 2001)

Instructions: This form has 34 statements about how you have been OVER THE LAST WEEK. Please read each statement and think how often you felt that way last week. Then tick the box which is closest to this.

not at all	only occasionally	sometimes	often	most or all the time
0	1	2	3	4

Example of items:

1. **I have felt terribly alone and isolated**
10. **Talking to people has felt too much for me**
12. **I have been happy with the things I have done**
14. **I have felt like crying**
28. **Unwanted images or memories have been distressing me**

Another similar tool, the *Outcome Rating Scale* (ORS), was recently developed by Miller and Duncan (see Miller, Duncan & Hubble, 2005). The ORS is a visual scale consisting of four items covering three areas (personal well-being, relationships, social functioning) as well as overall sense of well-being. The huge advantage of this tool is that it takes less than a minute to fill it in. A tracking system similar to the one used with the OQ-45 is being developed.

Qualitative Methods Investigating Changes Achieved by Therapy

An example of a qualitative method investigating changes brought by therapy is the *Client Change Interview* (CCI) (Elliott, 2002a). This is a structured interview that tries to assess what changes the client noticed since the beginning of therapy and to what extent therapy is responsible for them. The interview also considers hindering and helpful aspects of the treatment. The advantage of methods like CCI is their sensitivity to the wider impact of therapy, including its negative aspects (for this argument, see McLeod, 2001a).

Questionnaires Assessing Clients' Satisfaction with Therapy

Seligman's (1995) report of a survey carried out by the magazine *Consumer Reports*, which investigated clients' reported satisfaction with different aspects of therapy, is an example of usefulness of the methods of looking at client satisfaction. Some examples

of methods used in satisfaction studies are provided in McLeod (2003). In these questionnaires, clients are usually asked to answer questions on relevant aspects of therapy, e.g. how satisfied they were with therapy in general, to what extent therapy helped them to address problems that brought them to therapy, how they perceived the competence of the psychotherapist, whether they would recommend this psychotherapy to their relatives, and so on (see e.g. Larsen et al., 1979).

Experts' Evaluation

Experts' evaluation of change usually employs rating scales that are based on a structured interview. It is important that the expert who uses the rating scale is 'blind' to whether the person interviewed is in any kind of treatment. Otherwise the expert's judgement could be potentially biased. There exist several commonly used expert-rated instruments (see an overview in Hersen, 2004).

One frequently used instrument in the studies of depression is the *Hamilton Rating Scale for Depression* (HRSD) (Hamilton, 1960). The scale is used by the clinician for the evaluation of different aspects of depression (e.g. depressed mood, feelings of guilt, insomnia, psychomotor retardation, agitation, somatic anxiety). There are plenty of other rating scales for a variety of psychiatric diagnoses, for example the *Yale-Brown Obsessive-Compulsive Scale* (YBOCS) for obsessive-compulsive disorder (Goodman et al., 1989).

Another example of an expert-used instrument is the *Social and Occupational Functioning Assessment Scale* (SOFAS) that evaluates psychological, social and occupational functioning on a numerical continuum of mental health–illness (from 0 to 100). The SOFAS is based on Luborsky's *Health-Sickness Rating Scale* (Luborsky, 1962) that was revised by Spitzer et al. (Endicott, Spitzer et al., 1976; previously named the *Global Assessment Scale – GAS*). The SOFAS constitutes Axis V of *DSM–IV* (American Psychiatric Association, 2001).

For diagnostic as well as research purposes, especially in North America the *Structured Clinical Interview Diagnosis* (SCID) is widely used (First et al., 1996). It is a structured and standardized interview schedule aimed at improving reliability and validity in establishing *DSM–IV* diagnosis.

Therapist's Perspective on Outcome

One of the sources of the data on the evaluation of counselling and psychotherapy outcome can be the therapist. When thinking about the therapist as an evaluator of therapeutic change, we must not forget that he or she has a personal investment in the outcome of therapy. This makes the therapist's view somewhat dubious. The influence of this may have two aspects: in the direction of inflating therapy outcome as well as in the direction of deflating its outcome, probably depending on whether the therapist has a tendency to be rather self-critical or self-promoting.

One example of the scales that can be used by the therapist is the therapist's version of the Target Complaints scale. The therapist using this may judge the client's presenting problems and their intensity and then later in therapy repeat this evaluation.

Evaluation of Outcome by Significant Others

The evaluation of therapy outcome is sometimes performed by people close to clients in therapy, so-called significant others. Significant others are usually involved in rating the therapy outcome when the problem or the client population naturally calls for this kind of assessment. A good example, when it is definitely needed, is therapy with children, and to a certain extent therapy with adolescents. Typical significant others in that case are parents and sometimes educators or teachers who are in contact with the client. In therapy for adolescents (and in family therapy), there are often-used instruments, where adolescents rate their relationship with their parents, and vice versa (see e.g. Brent et al., 1997 and their study of therapy for adolescents with depression). Examples of measures that can be used as outcome measures in psychotherapy with children and youth are Achenbach's *Child Behavior Checklist* (CBCL) (Achenbach, 1999) and the *Youth Outcome Questionnaire* (Burlingame et al., 2001).

Another type of research that uses the reports of significant others are studies evaluating outcome of treatment for substance abuse (e.g. Babor & Del Boca, 2003). The obvious reason for involving significant others is the presence of denial in clients with substance abuse problem as well as the important fact that the clients with this difficulty negatively impact on their closest environment.

Marital and couple therapy logically also uses reports from significant others. Typically, the perception of own satisfaction and own view of the partner and the mutual relationship is measured. An example of a measure focusing on the behaviour of the other is a version of the Barrett-Lennard's *Relationship Inventory* (to be discussed further in Chapter 8). In this inventory, the partners mutually rate different aspects of their relationship, such as perceived empathy, positive regard, congruency and unconditionality of the other partner as well as self-perceived conveyed empathy, positive regard, congruency and unconditionality. Probably the most commonly used method in couple therapy outcome research is the *Dyadic Adjustment Scale* of Spanier (1976) that asks both partners about their perception of agreement on things like handling family finances, religious matters, sex relations or household tasks. The instrument also has a briefer form (Prouty, Markowski & Barnes, 2000). This and other couple and family measures are briefly presented in Jay Lebow's book, *Research for the Psychotherapist* (2006).

Despite the above examples, the evaluation of outcome by significant others is not common in current psychotherapy outcome research (McLeod, 2003) and is not without problems. The problem of ratings by significant others is that evaluation may be confounded by their lack of motivation or, on the other hand, bias in their relationship with the client. For example, using the rating of family members or teacher can be questionable as it can be affected by their own interests in relation to the client. For example, when

client changes result in negative consequences for the rater, e.g. the client becomes more assertive in school, the teacher may not always see it as progress.

Behavioural, Physiological and Other Objective Data

The nature of the problem being treated by counselling or psychotherapy sometimes requires its assessment by a behavioural, physiological or other objective measure. These may be somewhat conservative measures (see Lambert & Hill, 1994), for example monitoring physiological reaction (e.g. heart rate) to a fearful stimulus (e.g. spiders), in relation to specific phobias. In some cases behavioural data may be relevant, as in the case of agoraphobia (monitoring of the number of walks outside the house). Besides behavioural and physiological data, information on the length of hospitalization, frequency of absences from job, etc. can serve as objective data. An example of an objective measure is weight (Body Mass Index) in therapy for eating disorders (see e.g. Kächele et al., 2001). Similarly, in therapy for drug abuse urine check-ups are typical (Goldstein & Brown, 2003). Overall, this kind of measure is quite common for specific types of problem (those mentioned above) and more typical of behavioural therapy (see McLeod, 2003). More recently, the use of functional neuroimaging such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), or single photon emission computed tomography (SPECT) as tools for assessing the impact of psychotherapy is becoming more common (see Linden, 2006).

Evaluation of the Cost-effectiveness of Psychotherapy and Counselling

Recently, the monitoring of costs associated with counselling and psychotherapy has become more widely used. It enables a comparison of different therapeutic approaches and different types of treatment in regard to both costs and benefits – for example financial costs vs. improvement of symptoms or reduction of absence at work. For psychosomatic disorders, it may be a decrease in visits to the GP.

Estimating the cost-effectiveness and cost–benefits may sound a bit dehumanizing, but it seems that it is to a certain extent inevitable as the third-party payers responsible for public health want to spend money available for mental health wisely and justly.

Choice of an Outcome Measure

The first condition for the selection of an outcome measure for a concrete study is its reliability and sufficient information on its validity. It is probably wise to use established

measures as this enables a comparison of results from different studies. In regard to construct validity, it is sensible to use more measures (it is quite common to use at least four measures), to measure more variables simultaneously (e.g. self-esteem too in cases of depression), and to combine more perspectives (client, therapist, external rater, etc.). The most widely used outcome measures are the client-used self-report scales and expert-used rating scales. The researcher should also know what the magnitude of effect size he or she may expect with a particular measure, i.e. how sensitive the measure is. One should not forget that different measures may favour different therapeutic approaches. Some valuable information, though somewhat older, may be found in Lambert and Hill (1994: 83–84; see also Box 1.3.). Finally, a good overview of available methods is provided in the edited volumes of Maruish (2004).

Box 1.3 Lambert and Hill's (1994, pp. 83–84) conclusions about the sensitivity of outcome measures based on their review of outcome measures used in outcome research up to 1994

1. Data from therapists and expert judges, in which judges are aware of the treatment status of clients produce larger effect sizes than data from clients (self-report data), data produced by significant others, data from relevant institutions (e.g. employer, educational) and physiological data.
2. Gross ratings of change (e.g. whether or not the client has improved) produce a greater estimation of change than ratings on specific dimensions or symptoms.
3. Change measures based on the specific targets of therapy (e.g. specific symptoms) produce larger effect sizes than measures more distal from psychotherapy (including the common personality questionnaires).
4. Life adjustment measures that focus on social functioning in natural settings produce smaller effect sizes than analogue and laboratory based measures.
5. Data collected soon after therapy show larger effect sizes than data collected later.
6. Physiological measures (e.g. heart rate) usually produce small effect sizes compared to other measures even when they are specifically targeted in psychotherapy.