Consensus Group on Outcome Measures for Addiction
Acknowledgements

The Consensus Group on outcomes for addiction was funded through an NIHR Collaborations for Leadership in Applied Health Research and Care (CLAHRC) programme.

CLAHRCs undertake high-quality applied health research focused on the needs of patients and support the translation of research evidence into practice in the NHS. CLAHRCs are collaborative partnerships between a university and the surrounding NHS organisations, focused on improving patient outcomes through the conduct and application of applied health research. They create and embed approaches to research and its dissemination that are specifically designed to take account of the way that health care is increasingly delivered across sectors and a wide geographical area.

The Collaborations focus on the "second gap in translation" identified by Sir David Cooksey's (2006) Review of Health Research Review of UK Health Research; translating research into practice. They also address a key recommendation of the Chief Medical Officer's High Level Group on Clinical Effectiveness; the need for the NHS to harness better the capacity of higher education to support initiatives to enhance the effectiveness and efficiency of clinical care (Department of Health, 2007).

Nine CLAHRCs were established in October 2008, each with a number of themes. The Leeds/York collaboration (ARiAS) is the only addiction theme and one strand of its work is to build on the RESULT (Routine Evaluation of the Substance Use Ladder of Treatments) Project, a long-standing programme of collaborative research led by Leeds Addiction Unit and the University of York, School of Health Service Research.
**RESULT Project contributors**

Duncan Raistrick  
1980 – present: Joint Project Lead and Consultant Addiction Psychiatrist

Gillian Tober  
1984 – present: Joint Project Lead and Consultant Addiction Psychologist

Christine Godfrey  
1994 – present: Professor of Health Economics

Nick Heather  
1992 – 2007: Professor of Addiction Psychology

Steve Parrot  
1994 – present: Research Fellow in Health Economics

Helen Crosby  
2009 – present: Research Coordinator

Veronica Dale  
2011 – present: Statistician

Jenny Lang  
2009 – present: Research Practitioner

**Expert Group members – first convened March 2011**

Owen Bowden-Jones  
Consultant Addiction Psychiatrist

Alex Copello  
Professor of Addiction Psychology

Helen Crosby  
Research Coordinator

Veronica Dale  
Statistician

Ed Day  
Consultant Addiction Psychiatrist

Eilish Gilvarry  
Consultant Addiction Psychiatrist

Christine Godfrey  
Professor of Health Economics

Don Lavoie  
Department of Health

Damian Mitchell  
Department of Health

Duncan Raistrick  
Consultant Addiction Psychiatrist

Julia Sinclair  
Senior Lecturer in Psychiatry

John Strang  
Professor of Addiction Psychiatry

Gillian Tober  
Consultant Addiction Psychologist

Alex Whincup  
Specialist Midwife
Objective of the RESULT Project

The objective of the RESULT Project is to identify outcome measures suitable for use by specialist services. There is an abundance of possible outcome measures for both addictive behaviours and related problems (Deady, 2009; http://lib.adai.washington.edu). A rational approach to selecting from these is to apply a number of filters based on reasoned principles. Measures can be biomedical, for example liver function tests or birth weight, or they can be based on service user ratings, for example ratings of depression or dependence – these are collectively known as Quality of Life (QoL) measures.

Different reasons for using outcome measures

Different stakeholders want to know different things from the outcome measures and this needs to be acknowledged and accommodated when measures are selected. Stakeholders include service users, therapists, funders, researchers, public health teams and managers. To achieve the most direct QoL measure, self rating tools are the gold standard but there are circumstances where proxy measures are required: For example, if a service user is unable to read or understand a questionnaire, or if the service user rating might be biased as in child protection or criminal justice assessments.

Fitzpatrick et al. (1998) have described the advantages and disadvantages of different kinds of QoL measures and concluded that a judicious mix is likely to perform best and meet the varied demands made on outcome measures. Generic measures are aimed at painting a very broad picture of health or social status and so can be used to compare people with a particular condition, such as addiction, with people who do not have an addiction. The disadvantage is the loss of detail about the index problem. Utility measures are a kind of generic measure but have to do with preferences and costs and so have a utility for planning and prioritising health and social care. Dimension specific measures provide detail about an area of concern, addiction for example, but do not have the detail found in condition specific measures, dysfunctional families for example, which have the disadvantage that they can only be compared to the same condition specific measures. Finally individualised measures allow the service user to select matters of personal concern which cannot be compared to any other service user personal outcomes.

Four groups of measures have been adopted to reflect the different kinds of data described by Fitzpatrick et al. (1998). It is intended that measures should be used in a cumulative and hierarchical way starting with the most generic through to the most individualised outcomes. It is possible to use measures on a ‘pick and mix’ basis – however, too few measures will reduce the ability to make comparisons with other health care and addiction services and will give an incomplete picture. Equally too much data collection can distort clinical practice and become unethical.

Generic (including utility) measures
- Health status
- Health diagnoses

Dimension (addiction) measures
- Substance use and misuse
Dependence
Psychological wellbeing
Social wellbeing

Condition specific measures
- Mental illness – severe and enduring
- Mental illness – personality disorder
- Mental illness – neuroses
Families
Parenting and pregnancy

Individualised measures
- Physical health goals

Some principles of data collection
No single measure can adequately describe the variety of outcomes that can be described for any one person. RESULT integrates validated scales and other measures into a single package of options. Wherever appropriate measures are universal, which is to say that, as far as possible, they can be applied to any or all substances and are culturally and socio-economically neutral.

Although therapists will know their service users well the evidence is consistent in demonstrating that independent observers or service users themselves are the better source of both outcome prediction and outcome measurement – this does not mean that all therapist ratings are invalid rather that the greater reliance should be placed on the service user ratings (Tober et al., 2008).

Mandatory data collection should be avoided because the process of data collection itself becomes the goal rather than concentrating on collecting the highest quality of data. The RESULT measures are chosen as suitable for feedback at initial assessment. Outcome measures are best collected at 3 months after starting in treatment (this is when most change is expected) and then at 12 months (to test the sustainability of change) and thereafter annually (Raistrick et al., 2006; Babor et al., 2003; Weisner et al., 2003). Time frames vary from one measure to another and it makes sense to vary the original measures and standardise wherever possible.
Selection of recommended measures

The history of RESULT is the chronicle of a programme of research whose overall aim was to devise a method of measuring outcomes for service users that would be meaningful to service users, their friends, families and carers, the practitioners involved in their care, inspectors and commissioners of services. The programme was instigated by the Leeds Addiction Unit, working collaboratively with the University of York Department of Health Sciences, and the regional addiction services based in Newcastle upon Tyne.

Development of a measure of dependence

The project commenced with the development of a measure of dependence which, in the early 1980s would be consistent with contemporary theorising about dependence, namely that dependence is a psychological state with physical antecedents and consequences which are a product of substance use. Measures of the time were tied to patterns of use of specific substances.

Substance specific dependence scales that relied on a theoretical understanding of dependence in which tolerance and withdrawal were at its core needed to pinpoint specific withdrawal symptoms. In contrast a theoretical paradigm that identifies dependence as a set of thoughts and behaviours which are driven by the use of the substance and its consequences does not require the identification and measurement of specific withdrawal phenomena and enables the development of a measure that is not tied to a specific substance. Moreover, where addiction services are treating multiple substance use disorders, substance specific measures are going to cause confusion. A single, universal measure is an advance which allows comparison of levels of dependence across substances so that the severity of dependence can be tracked in individuals as they migrate from the use of one substance to another, or simultaneously use several. The Leeds Dependence Questionnaire (LDQ) was developed with these particular considerations in mind (Raistrick et al., 1994).

Discussion of the treatment goal stems from several aspects of dependence and other data. Measures of dependence that are tied to substance use, and specifically tolerance and withdrawal, are, by definition, not able to measure dependence in people who are abstinent. It has long been recognised that a state of dependence endures beyond the time limits of active substance use and is indeed responsible for the common phenomenon of relapse. Clinical decision-making needs information about dependence that is derived both during active use or a period of abstinence.

Clarity, face validity and brevity are pre-requisites to service user satisfaction and accurate information gathering. With this principle in mind, the proposition that dependence might serve as a proxy measure for drug and alcohol use was explored. In the development of the LDQ a strong correlation between use and dependence was found in help seeking populations. However, the need to measure dependence in a period of abstinence in order to make the right clinical decisions and decide the correct treatment goals was the overriding imperative in retaining a separate measure of substance use (Tober, 2000). The LDQ remains the only universal measure of dependence which has been shown to measure dependence during periods of abstinence.
Psychological and social stability domains

The essential purpose of specialist addiction services is the treatment of substance dependence and related problems. It can be argued that the measurement of dependence and substance use are sufficient for determining the quality of treatment outcomes. However, psychological wellbeing and social stability are known to be important predictors of outcome (Moos & Moos, 2006; Adamson et al., 2009), and, furthermore, problems across these domains are common reasons for help seeking. Initially the General Health Questionnaire (GHQ) (Goldberg, 1972) was selected on the grounds that it was widely used and well supported by psychometric data. It was superseded by the Clinical Outcomes in Routine Evaluation (CORE) measure (Evans et al., 2002) which is very similar to the GHQ but is freely available and has broader based psychometric data.

The measurement of social stability was a separate challenge. The choice was between using a problems checklist such as the Alcohol Problems Questionnaire (APQ) (Drummond, 1990) or a social satisfaction approach such as Corney & Clare (1985). It was decided that a brief measure would also need to be a generic measure and so the Social Satisfaction Questionnaire (SSQ) (Raistrick, 2007) was developed from the Corney & Clare (1985) questionnaire. The SSQ was just eight items which is brief as compared to the available problems checklists that inevitably are long, 30 or more items, in order to capture the possible range of problems.

Strengthening the key domain measures

The three scales LDQ, CORE, and SSQ are thought to be a good representation of addiction – referred to as dimension group of measures. The scales have been shown to be independent of each other and have been tested in large samples across agencies (Heather et al., 2001). At the same time it was recognised that cost effectiveness of services would need to be part of routine outcomes measurement (Coyle et al., 1997) and so the EuroQol (EQ-5D) (Szende et al., 2007) and the International Classification of Diseases (ICD-10) codes were added to the RESULT package in order to make comparisons with other healthcare services — referred to as generic measures. The EQ-5D was selected on the grounds that the National Institute of Clinical Excellence (NICE) prefer it as the measure for calculating quality adjusted life years. ICD-10 was selected on the grounds that it is the means of reporting diagnostic categories in NHS statistics.

The Maudsley Addiction Profile (MAP) (Marsden et al., 1998) was developed as a brief but comprehensive assessment tool and the best items were incorporated into RESULT. Typically these items generate a frequency score for health risk and social functioning behaviours over the previous month. More recently Health of the Nation Outcome Scores (HoNOS) (Salvi et al., 2005) have been incorporated for the same reason as the MAP items.

The question of how to measure substance use remains a challenge. For service users and for the general public substance misuse is the most obvious outcome measure. Practitioners can get good quality data on frequency of use but less good data on quantity. In many respects frequency alone is good enough (Tober, 2000) and is the principal substance use measure of MAP. It is possible to collect quantity data for alcohol misuse in the form of units of alcohol. RESULT uses the World Health Organisation instruments AUDIT (Babor et al., 1989) and ASSIST (WHO
Scaled measures
Where scales are selected they must demonstrate high specification psychometric properties. For full details of psychometric properties refer to original journal articles, user manuals or websites. Fitzpatrick et al. (1998) have described a number of criteria by which an outcome measure should be considered for inclusion in a clinical trial:

- Is the content of the instrument appropriate to the questions which the clinical trial is intended to address (Appropriateness)
- Does the instrument produce results that are reproducible and internally consistent? (Reliability)
- Does the instrument measure what it claims to measure? (Validity)
- Does the instrument detect changes over time that matter to patients? (Responsiveness)
- How precise are the scores of the instruments? (Precision)
- How interpretable are the scores of the instrument? (Interpretability)
- Is the instrument acceptable to patients? (Acceptability)
- Is the instrument easy to administer and process (Feasibility)

Appropriateness
The content of the instrument should be suitable to answer the questions that are being addressed. In considering the appropriateness of an instrument, the service user population, the question being addressed and what is actual being measured should all be considered.

Reliability
Reliability is the extent to which responses to scale items are repeated when the scale is administered on more than one occasion and by different practitioners. This is usually reported as the correlation between test and re-test scores. It is important that the time between the two ratings is long enough so that the service user cannot remember the first rating but short enough that change will not have occurred; one to two weeks is a good separation.

Internal reliability
Any scale that measures a single construct should have a high internal reliability. This means that all of the items in the scale are measuring aspects of this single construct. Internal reliability is normally reported using Cronbach’s co-efficient alpha. An alpha in the range 0.75-0.85 is ideal – above this it may be that there is redundancy of items and below this there may be poorly performing items.

Validity
Validity is the extent to which a scale measures what it claims to measure. For example, do the items of the Leeds Dependence Questionnaire truly measure the concept of dependence? It is never possible to prove that a scale is valid, rather, repeated tests of validity in different circumstances make it more and more likely to be true. There are different ways of determining validity:
• Scales are said to have a high degree of *face validity* when visual inspection of the scale items suggests that those items are appropriate to the particular construct. None of the RESULT scales seek to conceal their purpose and, therefore, are considered to have high face validity.

• The correlation between a scale and a gold standard is a measure of *construct validity*. If there is no gold standard then a scale that measures something similar can be used.

• The ability to separate out different populations, typically the general population from a problem population, is known as *discriminant validity*. A different, but essential, kind of *discriminant validity* is the ability of an outcome measure to be responsive to change over time.

*Responsiveness*
Responsiveness is a measure of the association between the change in the observed score and any clinical change. An instrument should be sensitive to clinical changes and be able to detect the effects of any treatment. Responsiveness is often expressed as an effect size.

Floor and ceiling effects are limitations in the responsiveness of an instrument. If a substantial proportion of service users achieve the lowest score then floor effects are present likewise if a substantial proportion achieve the highest score the ceiling effects are present. This indicates that a measure is limited in representing the whole range of severity of the condition.

*Precision (specificity and sensitivity)*
These measures apply where a scale is being used to identify a particular problem – for example, depression. Specificity is the extent to which true negatives are identified and sensitivity is the extent to which true positives are identified – in both cases a rating of 1.0 would be 100% correct identification. It is usual to measure specificity and sensitivity by using a research standardised interview to establish whether a condition, say depression, is present or not and then to compare the performance of a rating scale against the research practitioner's rating.

*Interpretability*
Interpretability is concerned with how meaningful the scores are. The scores should be meaningful to both patient and practitioners. A service user score should indicate whether they are above, below or at the normal level. The comparison could be by the use of benchmarking or population norms. Population norms allow more meaningful comparisons as they can split by age, gender and even condition. What is normal to a 25 year old male could be different to a 65 years old female. However these norms are not available for all instruments. (Bent et al., 2009).

*Acceptability*
The acceptability of an instrument will normally be worked through in the development stages. Important aspects are the use of plain English, ensuring that items ask a single question, and having a suitable range of responses. One way of measuring acceptability of fully developed instruments is to record the frequency of unanswered questions.
Feasibility
The feasibility of an instrument includes the time and resources required to collect process and analyse the instrument. Some instruments can be completed independently by patients whereas other instruments may require trained staff to administer the instrument.

Unless otherwise stated use of any of the outcome measures referred to here are subject to any restrictions laid down by the originators – permission to copy or use any of these scales should be sought from the original authors or by reference to the website.

Where more than one measure meets the inclusion criteria account is taken of pragmatic considerations:
- Current use of tools and measures in clinical practice
- Number of independent publications supporting measures
- Use of measures in UK research
- Extent of general and other treatment population data available

Systematic reviews of Quality of Life (QoL) assessment tools (Lang et al., submitted) and screening tools (Mdege and Lang, submitted) have been carried out to check that selected measures remain appropriate. Two other recent reviews have also been examined (NIMHE, 2009; Baker, 2009).
A Quality Framework

The framework below for scoring scaled instruments is based on the above criteria and is adapted from the scheme developed by NIMHE (2009).

<table>
<thead>
<tr>
<th>Main area</th>
<th>Scoring system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence base (3 Characteristics)</td>
<td></td>
</tr>
<tr>
<td>• No evidence</td>
<td>0</td>
</tr>
<tr>
<td>• Published evidence by scale developer(s)</td>
<td>1</td>
</tr>
<tr>
<td>• Independent evidence</td>
<td>2</td>
</tr>
<tr>
<td>Number of validation publications*</td>
<td></td>
</tr>
<tr>
<td>Psychometric properties (6 Characteristics)</td>
<td></td>
</tr>
<tr>
<td>• Reliability</td>
<td>Yes = 1</td>
</tr>
<tr>
<td>• Validity</td>
<td>No = 0</td>
</tr>
<tr>
<td>• Sensitivity for measuring change</td>
<td>Max. points 6</td>
</tr>
<tr>
<td>• Multiple settings</td>
<td></td>
</tr>
<tr>
<td>• Multiple cultural groups</td>
<td></td>
</tr>
<tr>
<td>• Translated versions</td>
<td></td>
</tr>
<tr>
<td>Availability of tool/measure (2 Characteristics)</td>
<td></td>
</tr>
<tr>
<td>• No training</td>
<td>Yes = 1</td>
</tr>
<tr>
<td>• Freely available</td>
<td>No = 0</td>
</tr>
<tr>
<td>Practical issues/considerations (3 Characteristics)</td>
<td></td>
</tr>
<tr>
<td>• Easy to score</td>
<td>Yes = 1</td>
</tr>
<tr>
<td>• Easy to interpret scores</td>
<td>No = 0</td>
</tr>
<tr>
<td>• No software required</td>
<td>Max. points 3</td>
</tr>
<tr>
<td>Universality (1 Characteristic)</td>
<td></td>
</tr>
<tr>
<td>• Applicable to any substance</td>
<td>Yes = 2</td>
</tr>
<tr>
<td>• Mostly = 1</td>
<td>No = 0</td>
</tr>
<tr>
<td>Service user rating (3 Characteristics)</td>
<td></td>
</tr>
<tr>
<td>• Readability</td>
<td>Yes = 1</td>
</tr>
<tr>
<td>• Clarity of meaning</td>
<td>No = 0</td>
</tr>
<tr>
<td>• Importance of measure</td>
<td>Max 3 points</td>
</tr>
<tr>
<td>Average completion time also recorded</td>
<td></td>
</tr>
</tbody>
</table>

*Supporting publications*

A review of the literature involved searches conducted in Embase (1980 to 2010 week 24), Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) (1948 to 16 June 2010), and PsycINFO (1806 to June week 3, 2011) for each scale denoted in the Quality Assessment of Scales table. The literature searches aimed to gather journal articles detailing validation in substance misuse treatment for each of the identified instruments. Searches were limited to English language. No date limits were applied. The search strategy combined alcohol and drug terms with the instrument title and associated acronym.
## Quality Assessment of Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Evidence (0-2)</th>
<th>Number of validation publications</th>
<th>Psychometrics (0-6)</th>
<th>Availability (0-2)</th>
<th>Practicality (0-3)</th>
<th>Universality (0-2)</th>
<th>Service User mean score (max 3)</th>
<th>Average completion time</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D</td>
<td>2</td>
<td>tbc*</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>LDQ</td>
<td>2</td>
<td>tbc*</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>CORE-10</td>
<td>2</td>
<td>tbc*</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>SSQ</td>
<td>1</td>
<td>tbc*</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>APQ</td>
<td>2</td>
<td>tbc*</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>HoNOS</td>
<td>2</td>
<td>tbc*</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>IRS</td>
<td>2</td>
<td>tbc*</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>FMQ</td>
<td>No data</td>
<td>tbc*</td>
<td>No data</td>
<td>2</td>
<td>No data</td>
<td>2</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
<tr>
<td>PCQ</td>
<td>1</td>
<td>tbc*</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>tbc*</td>
<td>tbc*</td>
</tr>
</tbody>
</table>

*t to be confirmed*
**Generic and utility measures**

This level is concerned with the relationship of addiction treatment to other health treatments. These high level measures will have low sensitivity for the nuances of addictions outcomes but are needed to answer questions such as: Is treatment cost effective? How ill are people with addiction problems compared to other illness groups? How complex are people with addiction problems? What is the morbidity profile of people with addiction problems? **RESULT** proposes using the utility measure EQ-5D and recording diagnoses using an internationally agreed system:

**EuroQol (EQ-5D)**
[www.EuroQol.org](http://www.EuroQol.org)

Selected because it is very simple to complete, there are extensive international, normative data, and it is the preferred method of calculating QALYs compared to other general health scales ([Szende *et al.*, 2007]). **NICE** uses QALYs to determine the cost effectiveness of treatments hence EQ-5D is commonly used in outcomes research.

The EQ-5D is a standardised generic measure of health status. It is applicable to most health conditions and treatments. Generates scale and cardinal scores derived from five dimensions: each dimension is assessed by one question scored one to three: one = no problem; two = some problems; and three = extreme problems. There are, therefore, 243 possible health states. These scores have no arithmetic properties and should be used as a cardinal score, however, a weighted score can be calculated and used to estimate quality adjusted life years (QALYs). The EQ-5D includes a visual analogue scale scored 0-100 which gives a global impression of how people see their health at the time of rating. The EQ-5D is extremely simple to complete and is undemanding in terms of intellectual ability required or of time.

**International Classification of Diseases (ICD-10)**
[http://apps.who.int/classifications/apps/icd/icd10online/](http://apps.who.int/classifications/apps/icd/icd10online/)

The **International Classification of Diseases** now in version 10 (ICD-10) has been developed by the World Health Organisation and is internationally recognised as a categorical system to classify mental and physical health problems and a wide range of related social problems that impact on health. ICD-10 provides brief descriptions of disorders to guide clinicians in reaching a diagnosis. ICD-10 is based on an alpha numeric coding scheme with a single character followed by two numbers at the three character level of detail (A00-Z99); further detail is then available by decimal numeric sub divisions of up to two numbers. For example: F = mental illness; K = diseases of the digestive system. The next two digits represent major groupings: F20 = schizophrenia; K70 = alcoholic liver disease. No codes for any normal states exist: the absence of ICD codes is, therefore a good outcome. The substance use codes can be used as a categorical outcome measure.
Addiction dimension measures

This level is concerned with the dimension ‘addiction’. The key domains of this level are common across addiction problems. These measures taken together paint a good general picture and answer questions such as: How severe is the addiction problem? How difficult is treatment likely to be? How does one addiction service compare with another? Do problems remain after abstinence or moderation of substance use?

Substance use

Substance use is the most direct way of measuring outcome – most other outcome measures are, in some way, related to substance use. However, there is no very satisfactory way of summarising and presenting substance use. For all substances there are two important measures, namely, the frequency and the quantity of use. Combinations of these can be used to generate meaningful categories such as binge drinkers, social drinkers, or harmful drinkers. Frequency of injecting can also be recorded with some reliability. Assuming the validity of self-report, the two main elements of an individual’s pattern of substance use can be measured in a variety of ways:

Quantity

A simple measure of quantity can be useful to determine the extent of substance use and to monitor progress in tackling this. For alcohol the situation is made easier in that all alcoholic drinks can be converted into standard units and it is common for treatment outcome trials to report on $%\text{days abstinent (PDA)}$ and $\text{drinks per drinking day (DDA)}$. No such standardisation is possible for the range of other drugs and frequency of use is probably the most objective and reliable measure for routine practice. Quantity may have to be measured in grams, rocks, joints, tablets, or in financial terms (£) and measuring quantity usually requires the individual to estimate their average use over a particular time period: for example, ‘How many joints of cannabis do you smoke each day?’ or ‘How much money do you spend on cannabis each week?’.

This is a practical strategy in that it is relatively easy to frame questions to capture these data, but by doing so the clinician may miss a lot of other useful information. For example

i. By summarising drug use as the usual patterns over the last month, unusual patterns of use may be missed, for example bingeing on cocaine over the weekend period.

ii. The style of drug use may be important, for example drinking alcohol until intoxication.

iii. Different routes of consumption lead to differing effects of the person, for example the same quantity of heroin will have a more potent effect if injected as opposed to being smoked or ‘chased on the foil’.

iv. A fundamental problem with illicit drugs is the relationship between the reported quantity of use and the amount actually consumed, as the purity of such drugs is usually unknown.

v. It is important to take into consideration that the reported amount of substance may have been shared with one or more other people.
Frequency

Frequency of use is often correlated with other problems, and is easier to measure than quantity. The simplest method is to present the individual with an ordinal ranking of frequencies of use and asking them to select the one that describes their situation in the time period under test, for example the last 28 days. However, the categories chosen, ‘daily’, ‘weekly’, ‘monthly’, will determine the accuracy of this method, and care needs to be taken to ensure that the person understands the meaning of the terms used, for example the difference between ‘often’ and ‘rarely’. An alternative approach is to ask the individual how many times they have used a particular drug in a set time period, the past week or past month and on how many days did they use. The longer the time period, the more difficult it will be to remember the exact details of use, and so more sophisticated variants of this method use a calendar filled in with important dates and events that are meaningful to the person to structure the process of recall. This is called the Timeline Followback (TLFB) technique and, although it improves accuracy of recall, it takes much longer and requires greater cooperation from the subject (Sobell et al., 1988).

Therefore, a hierarchy of assessment tools can be constructed that aim to capture different aspects of quantity and frequency of substance use:

1. **Current Frequency.** ASSIST (WHO ASSIST Working Group, 2002) is the simplest tool, and utilises a simple ordinal scale of frequency of use of seven categories of drug. The respondent is asked to select one of five categories of use over the preceding 3 months: ‘never’, ‘once or twice’, ‘weekly’, ‘monthly’ or ‘daily/almost daily’). The AUDIT-C (Bush et al., 1998) is validated for alcohol use.

2. **Current Frequency and Historical Use.** The Addiction Severity Index (ASI) (McLellan, 1980) drug and alcohol use section of the ASI uses the question ‘How many days in the past 30 have you used......’ and covers 10 different substances as well as alcohol. Further questions enquire about lifetime use and route of administration of each substance. It is therefore primarily a measure of frequency, although supplementary (unstructured) questions may be added about quantity and pattern of use.

3. **Current Quantity and Frequency.** The drug use section of the Opiate Treatment Index (OTI) (Darke, 1992) asks about the last three days of drug use for each drug category. The intervals between days of drug use, and the amounts consumed on these days, are employed to estimate recent consumption. Data are obtained on recent use in eleven drug categories (Darke, 1991). The OTI was designed to counter the problem of use of non-standardised amounts of substance by measuring use episodes, rather than the amount per occasion of use. This measure therefore reflects the degree of drug involvement of the individual, but although useful in research settings, the values obtained may have less utility in routine clinical settings.

More practical for routine use is the Maudsley Addiction Profile (MAP) (Marsden et al., 1998) which collects information about both frequency (number of days used in
the past 30) and quantity of use (amount used on a typical day). The Treatment Outcome Profile (National Treatment Agency for Substance Misuse, 2007) can be said to have evolved from the MAP, and uses a TLFB technique to estimate days of use in the preceding 4 weeks.

These different approaches reflect a spectrum of measurement of substance use that can be tailored to the treatment aims of a particular service. ‘Harm reduction’ goals may be evidenced using measures of quantity or frequency, whereas a simple dichotomous measure of use and non-use may be more useful for abstinence-orientated services.

Treatment Goal

**Abstinence**

**Prescribed drugs**

There is no established method for taking account of prescribed drugs as part of the whole outcomes picture. Clearly there is an outcome gulf between the person who is abstinent and not taking any medications at all and the person who is abstinent but prescribed, say, methadone 100mg and diazepam 40mg daily or the person who is abstinent but prescribed a naltrexone depot preparation.

The ICD-10 diagnostic categories deal with this quite well by having dependent use codes that include: Fx.21 abstinent but in a protected environment; Fx.22 taking a substitute drug; Fx.23 abstinent and taking a sensitizer or antagonist drug. A crosstab of the ICD-10 codes Fx.20 to Fx.28 creates a useful outcomes table.

In a medical setting at least all prescribing should be recorded and there should be a field available to confirm that a medication review has occurred at the follow-up period. There are established methods of undertaking such reviews (Murphy et al., 2009).

**Biomedical markers of substance use**

The fundamental question is ‘Has this person used the substance of interest in a particular time period?’ Laboratory toxicology screening is considered the gold standard for answering this question. However, this has practical, financial, and sometimes ethical implications. There are a variety of methods of testing for substance use (urine, saliva, sweat, hair, breath), and all capture a different time window of substance use. Interpretation of findings requires a knowledge of the screening techniques and the
potential for false positive or false negative results. Great care must be exercised if these measures are to be used for criminal justice or child protection proceedings.

Reviews of research on drug testing have concluded that there is good concordance (63-92%) between urine testing and self-report of drug use (Darke, 1998). However, this depends on both the substance and the population under study. For example, asking an individual new to drug treatment services about their illicit drug use is likely to yield different results than if the same questions were asked in the context of a long-term, supportive therapeutic relationship. Mathematical models of drug use patterns and drug screening schedules show that infrequent drug use is difficult to detect regardless of testing frequency, and that the benefits of more frequent drug testing are greatest with moderate levels of drug use (see table below).

<table>
<thead>
<tr>
<th>Drug Use Urinalysis Frequency</th>
<th>2x/wk</th>
<th>1x/wk</th>
<th>2x/mo</th>
<th>1x/mo</th>
<th>8x/yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Every day</td>
<td>3±2</td>
<td>7±2</td>
<td>15±10</td>
<td>30±13</td>
<td>46±40</td>
</tr>
<tr>
<td>Every other day</td>
<td>4±3</td>
<td>8±3</td>
<td>18±12</td>
<td>35±17</td>
<td>51±50</td>
</tr>
<tr>
<td>2x/wk</td>
<td>5±4</td>
<td>11±7</td>
<td>23±18</td>
<td>48±31</td>
<td>71±66</td>
</tr>
<tr>
<td>1x/wk</td>
<td>9±9</td>
<td>18±14</td>
<td>40±33</td>
<td>80±64</td>
<td>118±106</td>
</tr>
<tr>
<td>2x/mo</td>
<td>19±21</td>
<td>39±35</td>
<td>91±88</td>
<td>160±124</td>
<td>272±260</td>
</tr>
<tr>
<td>1x/mo</td>
<td>36±42</td>
<td>71±66</td>
<td>150±141</td>
<td>306±283</td>
<td>560±598</td>
</tr>
</tbody>
</table>


Biological markers are available as indicators of alcohol consumption. These markers have the advantage of being objective but are neither very specific nor sensitive to alcohol use. The search for better markers continues; there are a number that merit routine use (Whitfield, 2001; Conigrave, 2003): i) blood or breath alcohol concentration; ii) mean corpuscular volume (MCV); iii) serum gamma glutamyltransferase (GGT); iv) aspartate aminotransferase (AST); v) alanine aminotransferase (ALT); vi) carbohydrate deficient transferring (CDT); vii) HDL-cholesterol; and viii) uric acid. As with toxicology screening interpretation requires some medical knowledge.

**Alcohol Use Disorders Identification Test (AUDIT)**
The AUDIT was developed by the World Health Organization (WHO) specifically for use in primary healthcare (Babor et al., 1989; Saunders et al., 1993), but is now used in a range of settings. The AUDIT consists of ten items: three questions on alcohol consumption, four on alcohol-related problems and adverse reactions, and three on dependence symptoms. Apart from the last two questions, items refer to drinking in the
previous year and responses are weighted 0–4, based mainly on frequency of occurrence. Cutpoints on the AUDIT have been proposed as follows: >=8 hazardous drinking or worse; 8–15 indicates the need for simple brief interventions; 16–19 indicates the need for the addition of extended brief interventions; 20 or above indicates the need for referral to a specialist service (Babor et al., 2001). It is recommended that clinical judgement be exercised in cases where the AUDIT score is inconsistent with other evidence. The AUDIT-C uses the first three questions, alcohol consumption, and is a logical pairing with the Leeds Dependence Questionnaire. The National Institute of Clinical Excellence (NICE) guidelines on management of alcohol misuse and dependence (2011) propose the AUDIT as a simple and validated assessment of alcohol consumption.

**Alcohol, Smoking and Substance Involvement Screening Test (ASSIST)**

ASSIST was developed principally for use in primary health care settings consists of 7 questions covering tobacco, alcohol, cannabis, cocaine, amphetamine-like stimulants, inhalants, sedatives, hallucinogens, that could be answered by most subjects in around 10 minutes. A risk score is provided for each substance group. The **RESULT Project** has modified ASSIST to exclude alcohol and smoking. Cutoffs are: <=3 no intervention; 4-26 brief intervention; >=27 specialist treatment.

**Key Addiction Domains**

Central to the dimension of addiction are three independent constructs: i) dependence; ii) psychological wellbeing; and iii) social wellbeing. All three constructs may be both a cause and an effect of substance misuse. In many ways these measures paint a better picture of addictive behaviour than substance use itself and they are easier to measure but less intuitive to interpret.

**Leeds Dependence Questionnaire (LDQ)**
www.leedspft.nhs.uk/our_services/leeds_addiction_unit/RESULT

Selected because it can be used with any substance or addictive behaviour, does not depend on the presence of withdrawal symptoms and so measures dependence during periods of abstinence, has been used in major research studies, and has good normative data.

The LDQ is derived from a psychological understanding of the nature of dependence. The LDQ is an indicator of how addicted a person is and, therefore, how difficult it will be to achieve a positive outcome. For help seeking populations the LDQ is a reasonable proxy for substance use, however, for people who are socially quite stable, employed and having functional families, heavy drinking or other drug use is less well correlated with dependence. There are 10 items scored 0-3. Cut offs are: <11 = low dependence; 11-22 = medium dependence; and >22 = high dependence.

**Clinical Outcomes in Routine Evaluation (CORE-10)**
www.coreims.co.uk

Selected because it is easy to complete, is commonly used in mental health services, and has extensive normative data from general and help seeking populations. Symptoms of mental illness and diagnosis of mild to moderate mental illness is common among people attending specialist addiction services – CORE-10 can be used as an
outcome measure services specifically treating these conditions. In other words it is both a dimension and condition specific measure.

CORE-10 is a measure of psychological wellbeing. Although the scale measures a single construct, psychological wellbeing, inspection of individual items is an important component of risk assessment. There are 10 items scored 0-4 except items 2 and 3 are scored 4-0. Cut offs are: <10 = healthy or low level mental health problems; 10-25 = mild to moderate; and >25 = severe.

**Social Satisfaction Questionnaire (SSQ)**
http://www.leedspft.nhs.uk/our_services/leeds_addiction_unit/RESULT

Selected in preference to problem checklists because it is brief and has universality. As an outcome ‘satisfaction’ is more important to the individual than absolute measures of social function whereas for society the opposite is true. Normative data are available.

The SSQ measures social satisfaction which is taken as a proxy for social functioning and social problems. The latter two concepts are problematic in that ratings are highly dependent on socio-economic status, social roles and supporting social networks. In contrast social satisfaction is a more universal measure. There are 8 items which are scored 0-3 where a higher score indicates greater social satisfaction – there are no cut offs.

Data are available on the independence of the three key domain measures which taken together offer a useful profile of the effectiveness of substance misuse treatments. The three scales have been shown to be independent of each other and account for 71.3% of the outcome variance (see factor analysis below).

<table>
<thead>
<tr>
<th>Principal Components Analysis of Key Domain Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>component</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>LDQ1</td>
</tr>
<tr>
<td>LDQ2</td>
</tr>
<tr>
<td>LDQ5</td>
</tr>
<tr>
<td>LDQ7</td>
</tr>
</tbody>
</table>

26.1% of variance 25.4% of variance 19.8% of variance
Alcohol Problems Questionnaire (APQ)

Selected as an alternative to social satisfaction for services wishing a measure of actual problems. The scale is easy and quick to complete.

The APQ (Drummond, 1990) was developed for use in routine clinical practice and to be restricted to alcohol related problems (as opposed to dependence or use). The APQ is a 44-item questionnaire (maximum possible score of 44) which assesses eight problem domains: i) friends; ii) money; iii) police; iv) physical health; v) affective state; vi) marital; vii) children; and viii) work. The first five domains make up 23 items that are common to all individuals and a maximum score of 23 is derived from these items to arrive at a common score for all individuals. The problems list can be reworked to apply to drug use other than alcohol.

Other addiction dimension measures

A number of HoNOS (see below for a description of the HoNOS scale) items can be used to add a practitioner rating element to the self completion measures described above. These items are all rated 0-4:

1. Problems with relationships
2. Problems with activities of daily living
3. Problems with living conditions
4. Problems with occupation and activities

A number of MAP items can be used to add a practitioner rating element to the self completion measures described above. These items are all rated as the number of days out of the previous 30 days:

1. Days injecting drugs
2. Times injected drugs
3. Times shared a needle and syringe
4. Contact with family and friends
5. Days in employment

The Straus Bacon Social Stability Index is a simple 4-point scale: scoring is 1 point for the presence (yes) of each of the following criteria: i) stable employment over the past 3 years; ii) stable residence for the past 2 years; iii) current residence with at least 1 other individual (not currently living alone); and iv) current marital (or cohabiting) relationship.
Condition specific measures

Condition specific measures refer to services or treatment modalities for particular comorbidities. These measures are all answering questions on the specific target of the intervention such as: How severe was the target condition? How do services for the target group compare with each other? How effective was treatment with regard to the target symptoms/condition? Treatment programmes to measure include:

Mental Health – severe and enduring illness
The phrase severe and enduring mental illness covers a range of disorders but is dominated by people suffering from schizophrenia and severe affective disorders. There are 21 Payment by Results clusters for mental illness including a dual diagnosis cluster. Typically people with severe and enduring mental illness require the expertise and resources of community mental health teams who should take the lead clinical role (Department of Health, 2002) – specialist addiction services are likely to work on a shared care basis with mental health teams. Weaver et al., (2004) found some 36% of people in drug services and 53% in alcohol services have at least one diagnosis in this category.

Health of the Nation Outcome Scores (HoNOS)
http://www.rcpsych.ac.uk/training/honos.aspx
HoNOS is an internationally recognised outcome measure developed by the Royal College of Psychiatrists Research Unit (CRU) to measure health and social functioning outcomes in mental health services. It is the most widely used, routine clinical outcome measure used by English Mental Health Services. Although training is required, it has been stringently tested for acceptability, usability, sensitivity, reliability and validity.

The instrument has twelve items measuring behaviour, impairment, symptoms and social functioning. The scale is completed after routine clinical assessment by trained staff. Each scale is rated in order from 1 to 12. Information is not coded on more than one scale except for item 10 which is an overall rating. The rating is made on the basis of all information available to the rater (whatever the source) and is based on the most severe problem that occurred during the period rated (usually the two weeks leading up to the point of rating). The HoNOS system is not a standardised clinical assessment and cannot be a substitute for one. Adding up the scores of all 12 scales may not be particularly informative, as they are so wide in their coverage. Marked improvements in one domain may be cancelled out by deterioration in another, and it looks as if nothing has changed. Instead, changes in individual scales are used to assess progress. There are adaptations in other specialist areas, such as forensics, learning disabilities, and acquired brain injury, and the scale has been translated in numerous languages.

Mental Health - personality disorder
DSM-IV defines personality disorder is an enduring pattern of inner experience and behaviour that deviates markedly from the expectations of the individual’s culture, is pervasive and inflexible, and is stable over time. There is no single measure of severity of disorder that is easy to use in clinical practice and that is sensitive to change. The concept of personality disorder embraces distinct categories each of which exists along a continuum of severity – there is an association between severity, psychopathology,
substance use, and social problems. Bowden-Jones (2004) identifies 53% of alcohol service users and 37% of drug service user had a personality disorder: clinically people with emotionally unstable disorder make up the largest group. Many of those who also misuse substances will by default or by virtue of an agreed care pathway be seen by specialist addiction services.

**Impulsivity Rating Scale (IRS)**

*No website identified*

Impulsivity is a transnosological construct, present in normal as well as pathological behaviours, that covers a wide range of actions that are poorly conceived, prematurely expressed, unduly risky, or inappropriate to the situation. The Impulsivity rating scale (IRS) selected in preference to personality inventories of other specific questionnaires, as it is brief, and rates ‘real life’ behaviour in the last few days. It has shown reasonable validity, reliability and sensitivity to change across a number of service user groups and healthy volunteers. It is a 7-item hetero-evaluation of impulsivity which is clinician rated and takes 10-15 minutes. Scores range from 0-21 (Lecrubier *et al*., 1995).

**Other personality disorder measures**

A number of HoNOS items can be used to add a practitioner rating element to the self completion measures described above. These items are all rated 0-4:

1. Overactive, aggressive, disruptive or agitated behaviour
2. Non accidental self injury

**Neurotic disorders**

Anxiety and depression are very common among service users of addiction agencies. Commonly anxiety and depression are symptoms of substance use or withdrawal or related to consequences of addiction. Weaver *et al*. (2004) estimated that in drug services 27% of individuals had severe depression, 40% mild depression, and 19% severe anxiety; the corresponding figures for alcohol services were 34%, 47%, and 32%. CORE-10 which is recommended as a dimension measure can also be used as a condition specific measure, alternatively the Patient Health Questionnaire has been widely used for depression and the Generalised Anxiety Disorder scale for anxiety.

**Patient Health Questionnaire (PHQ-9)**

*Free and widely available from many sources on the internet*

The Patient Health Questionnaire was selected because it is patient administered, brief and is simple and quick to use. It also has excellent psychometric properties and is sensitive to change. It is a multi-purpose instrument which can be used for screening, diagnosing, monitoring and measuring the severity of depression (Kroenke and Spitzer, 2002).

The patient completes the instrument based on symptoms experienced over the last 2 weeks. The instrument has 10 scales which measure motivation, hopelessness, sleep, energy, appetite, concentration, negative thinking, physical activity, occupational role and suicidality. Only the first 9 scales are scored. Scores range from 0-27 with a higher score reflecting greater severity. Cut offs are: <4 = none or minimal; 5-9 = mild; 10-14 = moderate; 15-19 = moderate/severe; and >20 = severe. Proposed treatment actions are given for each cut off, for example, active treatment with pharmacotherapy and/or psychotherapy is suggested for moderate/severe depression.
Generalised Anxiety Disorder Measure (GAD-7)  
Free and widely available from many sources on the internet  
GAD-7 has shown value in assessing the severity of Generalised Anxiety Disorder, Panic Disorder, Social Anxiety Disorder and Post-traumatic stress disorder. It was selected because it is brief, simple to use, has good psychometric properties and is sensitive to change. Patients are asked to complete the instrument based on symptoms experienced over the last 2 weeks. The instrument has 7 scales which include anxiety, worry, loss of control, difficulty relaxing, restlessness, irritability and fear. Individual items are scored 0-3, with an overall score ranging from 0-21. Cut offs are: <4 = normal; 5-9 = mild; 10-14 = moderate; and >15 = severe.

Using the threshold score of 10, the GAD-7 is reported to have a sensitivity of 89% and a specificity of 82% for general anxiety disorder. It is moderately good at screening the three other common anxiety disorders: i) Panic disorder – sensitivity 74%, specificity 81%; ii) Social anxiety disorder – sensitivity 72%, specificity 80%; iii) Post-traumatic stress disorder – sensitivity 66%, specificity 81% (Kroenke et al., 2007; Spritzer et al., 2006).

Dysfunctional families  
A number of methods have been used in research studies in the UK and other countries to assess the impact of substance misuse as perceived by a family member: Symptoms of psychological and physical health, ways that family members cope and the social support they receive. A range of questionnaires has been used in these research studies and the psychometric properties of most of them are reported fully in Orford et al. (2005) and in Toner and Velleman (2011). These measures are based on the stress-strain-coping-support model of understanding of the effects on family members of living with a relative affected by addiction problems (Orford et al., 2010a).

Family Member Questionnaire (FMQ)  
No website identified  
The FMQ is a 33 item tool which pools the best items chosen using two psychometric criteria: i) items which, in a factor analysis of baseline data, load relatively highly on the factor corresponding to a sub-scale; and ii) items which show significant change in studies which follow up family members after a period in which they had received an intervention (Orford et al., 2010b). The questionnaire allows analysis at three levels: i) a total family pressure score; ii) scores for each element of the Stress Strain Coping Support model; and iii) scores for the eleven sub-scales. This is a new instrument and psychometric data are not yet available.

Other family work measures  
A number of HoNOS items can be used to add a practitioner rating element to the self completion measures described above. These items are all rated 0-4:  
1. Problems with relationships  
2. Problems with activities of daily living  
3. Problems with living conditions  

Pregnancy and Parenting  
Specialist addiction services often have a care pathway specifically for pregnant women – typically the pathway includes antenatal care and post natal parenting support. It has
been estimated that 250-350,000 children have drug using parents and 1.3million children live with problem drinking parents (refs from ALEX W). Serious injury and deaths are frequently associated with substance misusing parents.

**Parenting Concerns Questionnaire (PCQ)**
http://www.plymouth.ac.uk/pages/view.asp?page=35786
Selected because it is comprehensive and designed in line with current child welfare policy. There are no scales identified specific to substance misuse.

The PCQ was designed to help practitioners structure their assessment of child welfare within the context of three broad domains: i) Child development; ii) Parenting capacity; and iii) Family and environmental factors (Sheppard, 2010). The PCQ guides practitioners through the Common Assessment Framework (CAF) in relation to the needs of children and families. The scale has 37 items each scored 0-2 with 0 = no problem, 1 = problem but not severe, 2 = severe problem. There are no cut offs. The maximum score is 74.

**Other pregnancy and parenting measures**
It is usual to collect some obstetric measures as an indicator of the overall effectiveness of treatment through pregnancy:
1. Antenatal attendance record
2. Severity of neonatal abstinence syndrome (Dryden *et al.*, 2009)
3. Gestational age at delivery
4. Birth weight
5. APGAR score
6. Outcome of pregnancy – live birth, intrauterine death, termination
7. Placement of child at 6mth and 12mth
8. Child protection arrangements

A number of HoNOS items can be used to add a practitioner rating element to the self completion measures described above. These items are all rated 0-4:
1. Problems with relationships
2. Problems with activities of daily living
3. Problems with living conditions
Individualised measures

At this level the identification of personal goals is to help motivate and monitor progress for an individual and there is no expectation of generalising the outcomes.

The measurement of physical health problems presents a unique challenge as there is no limit to the range of those problems, which arise from addiction disorders. Every system may be adversely affected by substance misuse of one sort or another, and disease may or may not be treatable, thus the decision was made early on to relegate the outcome measurement of these to the individual level.
Interpretation of outcome measures

Slicing the outcome variance
The outcome variance is made up of all those factors that determine the difference between the best and the worst outcomes. It is usually not possible to account for more than 60-70% of the outcome variance and often rather less than this. Clearly, estimates of the impact of different ingredients of treatment on outcome depend on what is being measured. For psychosocial interventions service user characteristics are by far the most important determinant of outcome (Costello, 1980) whereas for pharmacological interventions individual characteristics are much less important. The following is a rough guide:

<table>
<thead>
<tr>
<th>Factor</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service user characteristics</td>
<td>60-70%</td>
</tr>
<tr>
<td>Follow up</td>
<td>10-20%</td>
</tr>
<tr>
<td>Therapist characteristics</td>
<td>10-20%</td>
</tr>
<tr>
<td>Specific treatment</td>
<td>10-15%</td>
</tr>
</tbody>
</table>

One of the several reasons that different treatments appear to deliver similar results is that the effect of the specific treatment is relatively small, nonetheless, there is evidence that selection of the right treatment for the right person at the right time is important (Witkiewitz et al., 2010) and may attenuate some of the adverse service user characteristics. In addition to this the quality of treatment delivery and the follow up and aftercare arrangements have all been shown to have a significant influence on outcome.

Predicting outcome at an individual level is notoriously uncertain, however, at an aggregate level psychological stability and social stability have consistently been shown to predict good outcome. Motivation, which should be assessed at the start of treatment, is important – those individuals who are highly motivated to change and who have the social and psychological resources to do so will probably get better without or in spite of treatment.

A general public, service user, and carer perspective
It is likely that the general public, service users, and carers will see abstinence from excessive drinking or illicit drug use as the only good outcome. Part of the CLAHRC study of outcomes for addiction includes a survey of service user views and this will be reported when the results are available.

Clinically significant change
The use of change scores, for example the difference between the mean LDQ score pre-treatment and post treatment, is not an adequate method of determining significant change because it fails to take account of the difference between individuals in the scores with which they start out and finish up nor does it take account the reliability of the instrument used to measure the change. Jacobson et al. (1999) proposed two criteria for asserting that clinically significant change, based upon the use of change scores, could be said to have occurred. These were:

- that the magnitude of change had to be statistically reliable
- that individuals end up in a range that renders them indistinguishable from well functioning people.
The calculation of statistically reliable change takes account of the reliability of the measurement instrument and thus avoids the pitfall of interpreting measurement error as change (Jacobson and Truax, 1991). Three methods for fulfilling the second of their criteria for clinically significant change, which have varying degrees of stringency, are proposed for asserting that change from the dysfunctional to the functional has occurred (Jacobson et al., 1999):

- scores that fall outside the range of the dysfunctional behaviour where this is described as being two standard deviations in the direction of improvement
- scores fall within the functional range where this is set at two standard deviations above the score for the normal population
- scores indicating the level of functioning suggest the service user is statistically more likely to be in the functional than the dysfunctional population

Bauer et al. (2004) compared alternative methods of estimating clinical significance and found that different statistical methods did produce different results. However, they concluded that the traditional Jacobson method was most suitable for estimating clinically significant change in clinical practice. So, taking the LDQ as an example, reliable change, RC, is calculated using the formula:

\[
RC = \frac{\text{Mean}_{LDQ10} - \text{Mean}_{LDQ13}}{S_{diff}}
\]

Typically this would be mean LDQ pre-treatment scores, LDQ10, and mean scores at follow-up at say 3mth., LDQ13. \(S_{diff}\) is the standard error of difference between the two means of LDQ scores, \(S_e\) is the standard error of measurement of LDQ, and \(S_1\) is the standard deviation of mean1. A reliable change score >1.28 is unlikely p<0.05 without actual change (Wise, 2004).

\[
S_{diff} = \sqrt{2(S_e)^2} \quad S_e = S_1\sqrt{1-r_{retest}}
\]

If a pre-treatment score is below the cut off point for ‘well functioning people’ then only reliable change, not clinically significant change, can be achieved. For this reason these individuals should be excluded from estimates of clinically significant change. In a study of the nature of dependence Tober (2000) reported on LDQ outcomes for a specialist service and found at 3mth. follow-up that 52.5% heroin and 48.7% alcohol users had achieved reliable change, and 33.9% and 28.2% respectively achieved clinically significant change. For CORE-10, Connell and Barkham (2007) report on outcomes for all participants and those above the cut off in a primary care counselling programme: 54.8% and 63.4% respectively showed reliable change with clinically significant change; 21.2% and 21.1% reliable improvement only; 21.6% and 13.7% no reliable change; 2.4% and 1.7% reliable deterioration.
Payment by Results

The Department of Health are currently undertaking pilot work to implement Payment by Results (PbR) across addiction services. PbR is a key Government policy describing funding based on outcome measures. Separate working groups are developing PbR clusters for alcohol and for drugs.

HoNOS has been developed to support mental health services and it makes sense to use HoNOS items in addiction services where they complement other addictions measures. HoNOS is routinely used in mental health services and recorded as part of the minimum data set (MHMDS) – mandatory data collection since 2003. The data set covers care received during an episode, details of clinical problems, treatment given, aspects of social care and outcomes (HoNOS scores).

The Mental Health Clustering Tool (MHCT) is based on the HoNOS and SARN scales. As previously stated HoNOS is an internationally recognised outcome measure developed by the Royal College of Psychiatrists Research Unit to measure health and social functioning outcomes in mental health services. The aim of the HoNOS was to produce a brief measure capable of being completed routinely by clinicians and recorded as part of a minimum mental health dataset. In contrast, the SARN scale was developed by the Care Pathways and Packages Project to aid the process of establishing a classification of service users based on their needs so that appropriate service responses could be developed both at the individual and service level.

The table below summarises the target outcomes for the scale measures recommended in RESULT. Clusters for alcohol, but not drugs, are available. The next phase of RESULT work will include estimating the proportion of people expected to achieve the target outcomes.
<table>
<thead>
<tr>
<th></th>
<th>SU Cluster 1 harmful/mild dependence</th>
<th>SU Cluster 2 moderate dependence</th>
<th>SU Cluster 3 severe dependence</th>
<th>SU Cluster 4 moderate/severe dependence and complexity</th>
<th>MI Cluster 16 Dual Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EQ5D</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LDQ</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;12 based on general population mean of 2.82 sd +/- 4.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CORE-10</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;10 based on cut off for healthy or mild mental health problems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SSQ</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;8 based on general population mean of 18.22 sd +/- 4.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>APQ</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HoNOS</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IRS</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td>Cut off ≥ 8</td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PHQ-9</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Depression rating: 1-4 minimal; 5-9 mild; 10-14 moderate; 15-19 moderate/severe; 20-27 severe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GAD-7</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anxiety rating: 5-9 mild; 10-14 moderate; &gt;15 severe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FMQ</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No cut offs. Maximum score 93; higher score = greater severity; 11 subscales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PCQ</strong></td>
<td>Target Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectation</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
<td>tbc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No cut offs. Maximum score 74; higher score = greater severity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Target scores are the optimum outcomes. Expected scores are the % in each cluster likely to achieve the target outcomes.
Development work in progress
RESULT will continue to evolve in the light of new developments and new data. Currently there are four strands of development work:

1. **Service user and public involvement.** Service users are being surveyed to gather views on the importance of different outcome measures. Service users are also being asked to rate the quality of the measures recommended in RESULT.

2. **Condition specific measures.** Scales and other measures are being identified that are specific to particular conditions. Currently the scope is limited to services for: i) severe and enduring mental illness; ii) personality disorder; iii) neurotic disorders; iv) dysfunctional families; and v) pregnancy and parenting.

3. **Combinations of measures.** Data are being collected to see how well the recommended measures work in combination. There will be statistical modelling to suggest how best to generate outputs.

4. **Payment by Results (PbR).** The outcomes will be mapped onto PbR clusters to indicate both the target outcomes and the proportion of service users in different clusters likely to achieve these outcomes. At the time of writing alcohol misuse clusters but not drug misuse clusters had been defined.
References


