Outcomes Tools

A Brief Guide on Measuring Outcomes

Phil Harris

Introduction

In recent years there has been increasing attention to the outcomes of what Government funded initiatives actually achieve. This has seen a shift in funding arrangements. In the 1980-1990's voluntary sector agencies were funded through grants. These were based on individual applications for local authorities to fund particular 'posts' to meet an emergent or established need. As such, they tended to be awarded on an ad hoc basis rather than as part of an overall strategic plan. These awards had no quality control in terms of set outcomes. Follow up studies subsequently revealed that many of these initiatives were not only ineffective but could actually make some ofthe problems they attempted to address even worse, famously in the case of car-mechanics courses for joy riders.

Changes to both UK and European law radically altered this funding structure in terms of introducing output measures attached to strategically commissioned services. The output measures, or Key Performance Indicators, set out the work programmes that organisations were expected to deliver. As such, output measures describe what the *organisation produces*. However, under the coalition Government there has been greater policy emphasis on outcomes. Outcomes refer to how the service effects change in *individuals who use the service*.

To facilitate the shift from output to outcomes requires the accurate measurement of change, its significance and to be able present these statistical results in a comprehensible format that is easily understood by a range of stakeholders. Furthermore, the commissioning cycle itself demands the collection, interpretation and strategic response to emergent data. In order to do this effectively, data 'noise' must be reduced in order to identify the most *relevant* data. Quality commissioning cannot be achieved where the data is inconsistent, unreliable or incomprehensible. Poor outcome data results in an inability to assess the performance of contracted organisations and destabilises effective future strategic developments.

A difficulty however, lies in the fact that these outcomes have been introduced rapidly and in a vacuum of knowledge. Measuring outcomes is a far more difficult than measuring an output. It is easy to identify how many clients were offered a counselling contract in any given period. It is far more difficult to measure what actually happened to the clients as a result of entering this service.

One approach to this problem is to adopt home grown tools. Whilst 'common sense' approaches have an intuitive feel, the purpose of statistical analysis is to go beyond the naïve assumptions that human beings often carry. For example, the use of leeches was common sense to doctors up until the Victorian period. Statistical analysis demonstrated that it was actually killing patients. Human beings are prone to bias, pet theories, assumptions and ideals that are never tested or systematically examined. For example, a believer in homeopathy may be convinced of their medication because it cured their cold in seven days. A 'believer' may be convinced by a stage-psychic who got three out of ten statements about them correct. A practitioner may believe they are measuring actual changes in alcohol use on a 10 point scale because the number changed. All these examples appeal to the observer because part of the experience pander's to their pre-existing assumptions-their common sense. But colds tend to last seven days. Probability states a psychic

will get half of their guesses correct. What does a change in score actually mean for the drinker? Understanding the validity, significance of these events requires a clear statistical analysis that goes beyond the immediate bias of the observer.
Establishing measurements that achieve a high degree of confidence is required if outcomes are to be effectively measured. This process is referred to as validation. An outcome tool that has not been through a systematic process of validation is simply not an outcome tool. The validation process is a stringent statistical examination that establishes the relevance of the tool, the magnitude of changes it detects and the statistical significance of these overall changes. It should also provide a framework for interpreting the results of the tool in a meaningful way. This is required to ensure that an outcome tool is actually measuring the real world target event and that it is doing so consistently.

What is a Variable?

Understanding outcome tools and any other psychological research process requires a clear definition of the basic unit of research: the variable. A variable is a characteristic, behavior, or value that can change. Variables are used in psychology research to determine if changes to one thing result in changes to another. So for example in measuring outcomes, did the treatment the client receive lead to a behavioral change? In research the variable which is manipulated by the examiner is the *Independent Variable* e.g. the treatment. The research then measures its impact and this is the *Dependent Variable* e.g. the treatment outcome. In a perfect world, it would be possible to control the environment so that all other influences were removed from the research. In this way the study would just measure the impact of the independent variable on the dependent variable. We could then measure exactly what impact treatment had on the clients gain as all other influences were eliminated. However, this is simply not possible. As we cannot eliminate all other influences, studies will always be influenced by *Confounding Variables*. These are extraneous factors that can exert influence the outcome. Confounding variables operate in two areas:

Participant Variables: These individual variables are related to the characteristics of each participant. This might include differences in background, mental health problems, number of previous treatment episodes, substance misuse severity etc.
Situational Variables: These environmental variables that influence how participant responds. These might include extra-therapeutic factors that are occurring in an individual's life, an inhospitable room etc.

In experimental conditions, these confounding variables are controlled for as much as possible. For example, a randomised control trial attempts to measure the impact of a specific therapy on outcome. It therefore must select participants that have similar clinical profiles or take a stratified sample to ensure that these factors do not interfere with the results. They must manaualise the approach to ensure that the treatment is delivered as intended. They must video the sessions to ensure the worker sticks to the manual. They have the same worker deliver more than one type of therapy in the study to calculate alliance factors in outcome. All of these are attempts to reduce confounding variables. However, not all can be eliminated. For example, in a randomized control trial clients are told they are in a gold standard research study. Practitioner's outcomes are deeply scrutinized which can increase

their performance rate. Clients can be paid to attend assessment days. These confounding variables *can* all have an impact on the dependent variable (treatment outcome). This makes it difficult to determine if the results are due to the influence of the independent variable, the confounding variable or an interaction of the two.
Finally, we need to create an operational definition for the test variables. This requires careful consideration. For example, what is the experimenter manipulating? We must be clear as to what interventions we are concerned with (brief, CBT, CRA, etc.). If outcomes are not linked to modality we cannot assess their performance. Furthermore we must define what is being measured?. This requires a clear definition of the outcome that you are trying to measure. For example, is it consumption rate? Is it positive lifestyle change? Is it the alleviation of negative symptoms? Is it that the person does not seek help anymore? Using different definitions produces very different outcome patterns. For example, people's consumption rates tends to be jagged for at least two years post treatment, not being in treatment anymore is not a sign someone is doing well but could be that they are alienated, the negative symptoms of depression may be gone but does someone have a more fulfilled life as a result?

Outcome measures can fail because they are poorly defined. For instance, in the outcome tool for young people in England it asks young people to rate their 'happiness' and 'anxiety'. These are not clearly defined variables as people interpret these very differently. Is happiness love, money, freedom from pain? Furthermore, both measures cross over. It is difficult to be "happy anxious''. Therefore improvement in one domain will inevitably be linked to improvement in the other. These measures are asking the same thing twice. Any improvements will be amplified across these measures because you cannot be one without the other. Re-wording the same question in this way is called a 'bloated specific' as it 'bloats' the overall impact on the outcome score.

This gives rise to the issues of the weighting of variables. Not all variables have equal weight. For example, an outcome tool that measures 'knowledge of drugs and alcohol' as well as 'improvement in life' using similar scales offer equal statistical weight to a non-relevant and relevant measure. So someone who gained considerable knowledge but made no change achieves the same outcome as someone who made considerable change but gained no knowledge.

Deciding on variables is vital in outcome monitoring. This is because whatever variable is decided upon as the outcome measure, it will serve as the benchmark of the agency's performance. It is would be very unwise for an agency to ignore this fact. So, instead they will invest their resources in the achievement of the benchmark at the expense of other work. For example, if an agency's outcome is to stop people returning to treatment for 12 months, they are liable to hinder treatment re-entry in this time frame. This is not a devious response but just human nature, as witnessed in the failings of many 'target driven' health care services in the UK. In short, whatever outcome you set is what you will get, at the expense of anything that is not measured.

Validity & Reliability: Overview

So, at its most basic level, an outcome tool must have clarity regarding the variables it is controlling and the variables that it is measuring. It must then undergo stringent examination in order to establish whether these variables are measuring real world change and that they are doing so consistently. This is referred to as validity and reliability.

Validity is the degree that the test measurement is well-founded and corresponds accurately to the real world criterion. The validity of a tool relies solely on the degree to which it measures what it claims to measure. An outcome tool must also be sensitive to detecting changes in the client's life and that these changes can be attributed to the treatment that they have received. Reliability is the extent to which a measurement tool provides consistent results. This is to say it can consistently identify true positive cases and true negative cases.

Just because a measure is reliable, it is not necessarily valid and vice-versa. For example an outcome tool measuring changes to self-esteem levels in depressed individuals may produce reliable results that are consistent. But as self-esteem is not related to depression the results would not be valid. Likewise a broken watch is valid as it is measuring the passage of time. But it is not reliable because it keeps giving different times. Validity and reliability are based on degrees rather than an absolute measure, however, only where validity and reliability are high can an outcome tool be adjudged to be validated. For example, cross referencing validity and reliability demonstrates how outcome tools that have not been validated provide highly distorted outcomes.

	LOW VALIDITY	**HIGH VALIDITY**
LOW RELIABILITY	a. This outcome tool would fail to identify the real changes the client made but its inconsistent result might produce some false negative and positives by random chance.	b. This would identify the changes the client made but the magnitude of these changes would be grossly distorted.
HIGH RELIABILITY	c. This tool would produce a consistent pattern of outcomes but these would not reflect the client's real-life change at all. Outcomes produced would be beguiling due to their consistency but they are consistently wrong.	d. This tool would produce a stable pattern of outcomes that were reflective of the clients change. It would cluster 'true' positive and 'true' negative cases.

Establishing Validity

Establishing validity involves a number of processes to ensure that the tool is measuring the targeted real world construct. The starting point is to establish ***Face Validity. Face Validity*** refers to whether the measure *appears* to be assessing the intended variable under study. It is not a scientific type of validity, but rather the product of the investigator's belief in what needs to be measured. As such, it is not a rigorous process and still resides in the realm of common sense error. Similarly, ***Construct Validity*** is used to ensure that the tool is measuring what it is intended to measure (i.e. the construct), and not other variables. This is similar to face validity but is established through a panel of experts who can collectively decide on what variables should be measured. Experts draw up the items for inclusion in the outcome tool and ensure that they cover a representative sample of the behavior being measured.

This remains prone to the bias's and allegiances of the investigating team-*expert bias*, when the experts believe they are measuring what is relevant but are in fact not doing so. For example, a panel of experts designed a risk profile to measure the complexity of young people's needs in England. To do this, they used the number of services supporting a young person as an indicator. This included whether the young person had a mental health worker, substance misuse worker, social worker etc. As a result, they were actually measuring young people's involvement in services and not the complexity of their need. For example, a young person with complex needs without services supporting them would have scored 0. This would have serious repercussions as funding would be allocated to areas according to this score. This would have sent money to service "rich" areas and not where services were needed.

It is vitally important that the outcome tool measures the full breadth of the variables being treated. For example, CORE is a validated treatment outcome tool. It measures 36 variables but these relate almost exclusively on the abatement of negative symptoms of depression and anxiety. Though it is used widely, it will only measure variables relevant to those clients who experience depressive and anxious symptoms. Conversely, the risk profile tool for young people in England has no reference to mental illness at all now so important changes will not be detected either.

This problem of validity can also occur in terms of when samples are taken. If outcome measures are taken at treatment entry, and then 6 months later at treatment completion, they may not be measuring client outcomes. Rather they are measuring the 'outcomes of those who completed treatment.' So they are measuring a sub-population not the whole population. Real time feedback tools, that measure on-going change, eliminates this problem and shows superior outcomes as a result. Real-time feedback offers instant data on the client's progress as it occurs. When clients deviate from the expected response it can trigger a red flag in order to address this decline in their progress. This alerts the worker and allows for rapid intervention. As a result, validated real time outcome tools do not simply measure an outcome but improve an outcome at the same time. As such it gives whole population data and not just data on completers.

Criterion Validity establishes the correlation between the test and the variable being measured. It does this by comparing test results with the results of other outcomes tools which have already been proven to be valid. This means a new outcome tool must be tested alongside an established one in order to see if they produce a similar range of outcomes to each other-correlation. This can be done in one of two ways. If the new tool is trialed at the same time as the older tool, this is referred to as ***Concurrent Validity Evidence***. For example, outcomes as measured in the Beck's Depression Index should be similar to the changes in depression as measured in the CES-Depression scale when taken at the same time period. If the measurements of the new tool correlate with the establish tool, then it is more likely to be valid. This can also be demonstrated in a slightly different way. The new tool can be tested by taking pre-test scores that are then used to predict outcomes collected at a later point in time. This is called ***Predictive Validity Evidence***. For example, the Complexity Index (Revised) scores at intake are predictive of the hierarchy of outcomes as measured by ORS at a later stage. Those with early onset externalized scores were predicted as achieving a lower range of outcomes than late onset normative youth.

Statistical Conclusion Validity is the degree to which conclusions about the relationship among variables is 'reasonable'. As such it is more concerned with the methods by which the tools are established, such as were there adequate sampling procedures to ensure a wide cross section of people were included in the study? Were there appropriate statistical tests conducted? What procedures were in place to ensure reliable measurement? For example, TOPS has been used with problem drinkers but the sample was taken from street drug agencies, there was an inadequate sampling size to establish any statistical conclusions and the items on the TOPS questionnaire do not have Construct Validity regarding drinking. As such, TOPS fails to meet the statistical conclusion validity for problems drinkers.

Internal Validity is an estimate of the *causal* relationships between variables (e.g. cause and effect). This requires the study of how an independent variable influences a dependent variable in highly controlled conditions. So for example, if you are measuring an outcome, how do you know that the changes in the clients score are the result of the treatment that they are receiving and not as a result of some other influence? There can be confounding variables that are actually influencing this change. For example, situational depression tends to occur for three months and then remits anyway, was there a major drug bust between the first and second measurement meaning that clients are not using now anyway? Did the first measure on the outcome score influence the client's second score? Was the sample population characterized by abnormally high or low scores?

External validity concerns the extent to which the findings of an outcome tool can be generalized to a wider population outside of the sample group. If the same research study was conducted on those other cases, would it get the same results? A major factor in this is whether the study sample are representative of the general population along relevant dimensions. **Ecological validity** is whether the outcome tool can be applied to real life situations outside of test settings. To be ecologically valid, the methods, materials and setting of a study must approximate the real-life situation that is under investigation. It must also consider the skills, training, length of time of application that delivering it will require. For example, the Maudsley

Addiction Profile (MAP) can take 1-2 hours to complete with a client. This deep assessment may be useful in a research trial but not in everyday practice.

There can be a tension between internal and external validity. Internal validity requires the elimination of as many confounding variables as possible in a highly controlled environment in order to establish causal relations. However, external validity occurs in everyday environments which cannot control these variables to such a high degree. Staff training, no control groups, idiosyncratic delivery of treatment, the ability to revise an approach for a non-responsive client are all key features of the real world unaccounted for in controlled conditions. This is why it is difficult to achieve internal and external validity in control trials.

Diagnostic Validity is especially important in outcome tools. This refers to the degree that the outcome tool achieves its objective. For example, does a new screening tool identify 'true' positive cases or discriminate against 'true' negative cases? Does the outcome tool identify 'true' positive outcomes as well as 'true' negative outcomes? This means the tool needs to demonstrate a high degree of *sensitivity and specificity*. The tool needs to be sensitive enough to detect the relevant problem if it is present (true positives & true negatives) and but specific enough not to respond to other things (false positives and false negatives). This requires testing against other diagnostic criteria but there can be dangers here as well, when both diagnostic test share the core assumptions. So diagnostic criteria of the ICD-10 and the DSM IVTR show high consistency with each other in the diagnosis of substance misuse with young people. Any one comparing these tools might therefore feel confident in these constructs and design an outcome tool based on these findings. However, latent class analysis shows that both diagnostic classifications generate a high rate of both false positive and false negative cases because they both share the core assumptions leading to a systematic bias in their reporting. They generate a high rate of diagnostic imposters-young people who look like they have a substance related problem but do not and diagnostic orphans who have a problem which is not identified by either criteria.

Reliability

Without validity, no one can be confident in the tool being use. The outcome measure will be providing data, but the value of the data is wholly questionable. This means that commissioners will be making decisions on extraneous factors and not the target behavior. Reliability is the degree to which an assessment tool produces stable and consistent results and can be established in a number of ways.

Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time to a group of individuals. The scores from Time 1 and Time 2 can then be correlated in order to evaluate the stability of the test results over time. **Parallel forms reliability** is a measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same variable, skill, knowledge base, etc.) to the same group of individuals. The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions.

Inter-rater reliability is a measure of reliability used to assess the degree to which different judges or raters agree in their assessment decisions. Inter-rater reliability is useful because human observers will not necessarily interpret questions or answers the same way; raters may disagree as to how well certain responses or material demonstrate knowledge of the construct or skill being assessed. One commercially available outcome tool that was not validated at release, uses a 10 point scale to rate different demands in the client's life. Subsequent research found up to a seven point variation in response when the client was asked the same questions by two different workers. This means that there could be 70 per cent variation in the basic scores of client, even before any analysis has been conducted. This error will not be identified in the second analysis leading to highly misleading data. They tried to resolve this issue by stating a definition for every point on their scale. This might improve their reliability but at the expense of validity. Now the tool may not be measuring the client's change but is instead measuring how well the client fits their descriptions.

Internal consistency reliability is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results. **Average inter-item correlation** is a subtype of internal consistency reliability. It is obtained by taking all of the items on a test that probe the same construct (e.g., drug use), determining the correlation coefficient for each *pair* of items, and finally taking the average of all of these correlation coefficients. This final step yields the average inter-item correlation. **Split-half reliability** is another subtype of internal consistency reliability. The process of obtaining split-half reliability is begun by "splitting in half" all items of a test that are intended to probe the same area of knowledge in order to form two "sets" of items. The *entire* test is administered to a group of individuals, the total score for each "set" is computed, and finally the split-half reliability is obtained by determining the correlation between the two total "set" scores.

Interpreting Data

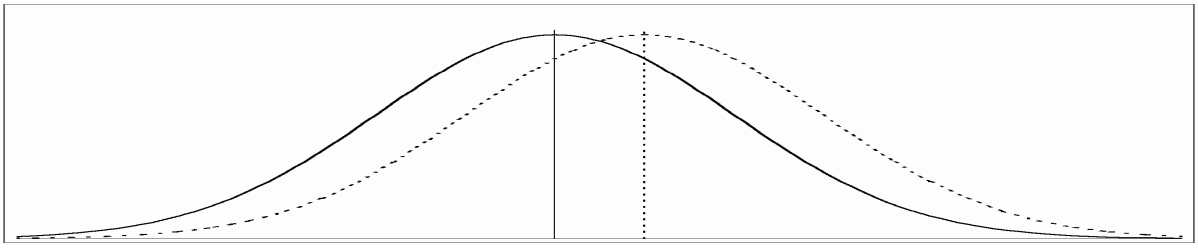
Once validity and reliability are known, there comes another problem of actually interpreting what the outcome is revealing. Outcome tools are usually based on the statistical analysis of raw data collected. This data must be translated into a readily understood format. Methods of statistical analysis are used to translate these findings into a picture of change. These include the statistical significance of change and the size of change.

Statistical Significance: As we saw in the examination of variables, change may be occurring for a number of reasons. Treatment, remissions, time, changes in drug trends all conspire in real world settings to influence a client's overall outcome. These confounding variables therefore need to be accounted for. This is done through the examination of statistical significance. Statistical significance describes how likely the changes which the tool has measured could have happened by random chance. The lower the probability that this was a random event in a treatment population then the more confident that people can be in these findings. This is set as a P-value, usually before the research findings are analysed. P-Values are often set along with an alpha level at 0.05. Statistical significance is achieved when scoring under this range as it very unlikely to have been caused by random

chance. Equations that establish statistical significance include the size of the sampled population in the formulae. This means that the size of the sample becomes very important and influences the final analysis. Statistical significance therefore requires a larger sample population in order to be reliable.

Effect Size: Ultimately an outcome tool must detect the magnitude of the change that has occurred. The easiest way of doing this is through **Effect Size**. This uses the standard deviations of normative distribution curves. What this means is that in a truly random sample of untreated clients, a few would have severe problems, most would have mid-range of problems and few would have modest problems. This is represented as a classic bell curve. Half this population is randomly selected and then treated. However, they will not all respond equally to treatment. Treatment will produce another bell curve. Some clients will have high response, most will have a middling response and a few will have a low response. Effect Size measures the difference between the mid-point of the non-treated group verses the mid-point of the treat group as shown below.

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$



Effect Size

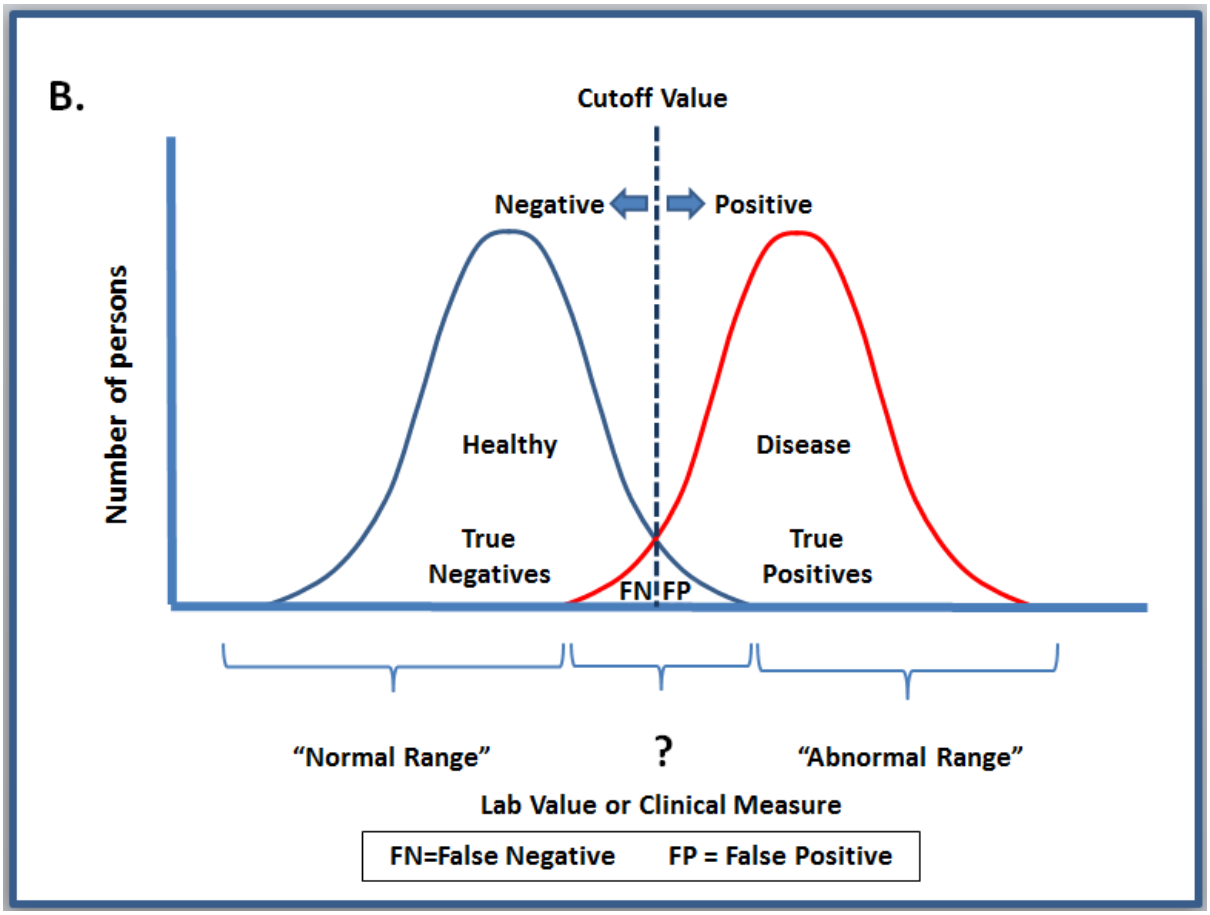
Effect size is reported on scale. An effect size of 0.0 strongly suggests that there was no difference in the average score of treatment and non-treated clients. Alternatively, an effect size of 1.0 would indicate that the worst performers in the treatment group were as a well off as the best performers in the untreated group. The size of this change is reported by Cohen's *d*. An effect size of 0.0-0.3 is considered weak, 0.4-0.7 is considerate moderate, whilst 0.8+ is considered strong. However, Cohen's *d* is very rudimentary measure. It is based on visually detectable difference. So, in the example of height, an effect size of 0.0-0.3 means that no differences between people could be observed. Between 0.4-0.7 differences could be noticed, whereas difference between 0.8+ would be obvious. However, this scale is not so apparent when applied to abstract constructs like lifestyle change limiting its applicability.

Effect Size measurements are very useful in comparison studies. Because it measures the differences between two curves it does not matter which tools were used to establish the curves. Therefore it tends to be used as in meta-analytic studies. This allows researchers to translate hundreds of studies into one large scale study, regardless of the measurement tools use. So, an outcome tool that can report effect size is useful for comparative purposes.

Clinical Cut Off Scores

Most data produced in statistics occurs on a spectrum of values from low to high. This data is revealing but sometimes it is necessary to know at what point on a spectrum a decision needs to be made. Statistics produce continuous spectrum of data but human decision making is binary. For example, a patient may produce a range of data on systolic measures but at what point does the surgeon decide to operate? On this spectrum there must be a point on this scale that separates normal from abnormal. In this case, abnormal is described as the point in which an individual would benefit from treatment. This is the ***Clinical Cut Off*** score. This also relates to specifically reviewed earlier. If a clinical cut off score is too high then many abnormal problems are neglected. If set too low then unnecessary treatment is given to people who do not need it. The same goes for the binary decision of understanding treatment outcomes. Clients scores may rise and fall but at what point do they cross over from abnormal to normal ranges of functioning? When is the client worsening?

Effective outcome tools do not simply produce a continuous range of values. An outcome tool needs to establish the cut-off points that identify what these values mean. Averages cannot be used for this. For example, national average weight changes over time and is not indicative of when diabetes type II will occur. Instead, clinical cut off scores are derived from the standard deviations of abnormal and normal bell curves. Essentially, it is the point when the standard deviation of the clinical sample intersects with the standard deviation of a non-clinical sample. This allows of the estimates the score at which a treated subject has a greater probability of belonging to the non-clinical sample. There are a number of ways of calculating this, but estimates come out roughly the same regardless of the complexity of the calculation used.



Clinical Cut-Off Scores

Reliability Change Index

As we have seen, reliability occurs in degrees so it is necessary to be able to compensate for this statistically. In order to establish whether the change in the clients outcome score is greater than could be expected by random chance, a reliability change index is used to calculate the cut off points of change. This approach calculates the degree of unreliability in scores (the margin of variation in clients answers when taken at two points in time) called the Reliability Index. If the client's reported improvement is greater than this margin of unreliability, then the patient can be said to have experienced clinically significant change. This is calculated as:

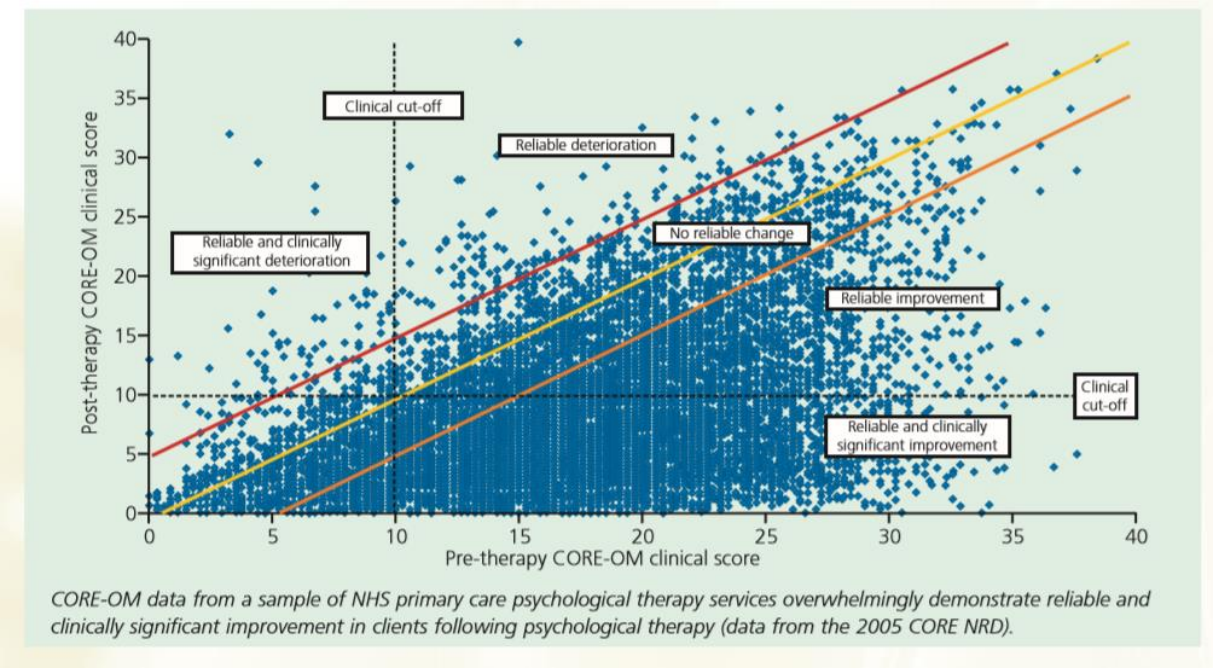
$$\text{Square root } (2\text{x Standard error of measurement})^2$$

Reliability Change Indication calculates a range of outcomes to offer binary meaning to the quality of change that has occurred for the treatment population. It uses the Reliability Index to do this. There are only four possible outcome from receiving psychosocial treatment. This includes:

- Deterioration: The clients change in score is lower than the range of unreliability
- No Change: The clients change in score occurs within the range of unreliability
- Reliable Change: The clients change in score is above the range of unreliability

- Clinically Significant Change: The clients change in score is above the range of unreliability and crosses the cut-off point of the normative population

Once the clinical thresholds are established from a sample population, the cut off thresholds for the quality of change can be determined. This means that outcomes can be reported across these four domains. For example, 100 clients were to receive counselling, their outcomes can be reported by what percentage deteriorated, experienced no change, reliable change or clinically significant change. This data is very important in light of the fact that there are very limited resources within the field to conduct long term follow-up of client's outcomes. What is known from research is that the client performance at the completion of treatment is highly indicative of their long term outcome. So those who are deteriorating in treatment tend fare badly at long term follow-up whilst those who have maximum benefit from treatment tend to do well. Therefore, the identification of the number of clients achieving clinically significant change is a cost-effective proxy measure of long term outcome. Furthermore, research also suggests that in the case of deterioration, practitioners and agencies are unlikely to respond to this unless there is threat to their personal or organisational reputation. Identification of negative cases therefore leads to greater concerted efforts to reduce this number.



CORE-OM data from a sample of NHS primary care psychological therapy services overwhelmingly demonstrate reliable and clinically significant improvement in clients following psychological therapy (data from the 2005 CORE NRD).

An example of the RCI for CORE

Even the use of RCI's can be further expanded. For example, if an organisation produces a RCI of its counselling outcomes, how do you know whether this is a good range of scores? This is where benchmarking becomes important. Most advanced outcome systems now utilize databases. Anyone using the tools can input their data in the database which will produce effect size and RCI reports. In return, it will also crunch this data from multiple users and create global averages of treatment outcomes for clients. This means that a practitioner, agency or region can see how their outcomes compare to this global performance. This cannot be based on averages. The average response rate is the mid-point in a range of scores and so does not offer insight into the performance levels of the wider population. Therefore,

these comparisons have to report against the RCI range of the wider treatment population.

This benchmarking is important in an age of outcome monitoring. In England, rather artificial outcome targets are being set, which are not based on clinical data but on figures which are thought of as optimal. This means that treatment outcome goals are naïve at best and impossible at worst. Benchmarking also offers greater assessment of agency performance in comparison to the global average of what can be achieved with this client cohort. Some of these systems, like FIT, have software that allows these data bases to integrate into existing client management programmes for ease of data translation.

Economic Viability

Another important consideration is economic viability. It is possible to develop home in-house outcome measures. However, this requires a high level of expertise, a large sample size of treatment and non-treatment groups, lengthy time periods between sampling and considerable analysis and development time. Without these measures, the development of a tool is severely compromised. This is economically expensive, where commercial tools are available at a significantly lower cost. Consideration would have to be given to whether the overall cost and maintenance of the 'in-house' tool would confer sufficient additional benefits over that of a commercially available tool to justify this level of investment. Furthermore, the development of 'in-house' tools would still require a large investment of time and development without a validation process. The worst case scenario is significant investment is made in an in-house tool that lacks validity or reliability.

Summary

All outcomes tools must meet minimum requirements in order to ensure confidence in the data reporting. Any suggested tools should be evaluated against the following criteria:

1) Measured variables must be relevant and applicable to the total client population
    - The outcome must be clearly defined
    - The variables measured must be applicable to this outcome
    - The chosen measure will influence agency priorities

2) The tools must demonstrate statistical validity and ensure it measures the targeted real world phenomena
    - Tools cannot be based purely on face or construct validity
    - Tools need correlation with other established outcome tools
    - Tools need to demonstrate concurrent or predictive validity

3) The tool must have been examined in terms of its reliability tin that it produces a consistent range of outcomes
    - Tool should show acceptable levels of test-retest correlation
    - Tool should have acceptable inter-rater reliability
4) It is vital that the tool has high diagnostic validity in terms of the identification of true positive cases and the elimination of true negative cases

5) Outcome measures that use real time feedback to enhance outcomes
6) Data from the outcome tool must be tested for its statistical significance in order to establish whether changes are by random chance or through the independent variable (i.e. the treatment) in order to establish confidence. This requires a large and representative sample size.
7) The simplest measure of the magnitude of change should be calculated by establishing the effect size for comparative purpose with other research findings
8) The tool should have established cut off points that identify the threshold of positive and negative cases. This requires a large treatment and non-treatment sample.
9) The tool should have a Reliability Change Indication in order to interpret the magnitude outcomes for the clients
10) The tool should be able to benchmark its outcomes against global averages of similar clients in order to compare real life performance
11) The tool should have an IT solution that allows for clear reporting
12) The adoption or development of a tool needs to be economically viable

Conclusion

The implementation of outcome measurement tools is central to the future development of commissioning. This requires commissioners and stakeholders have a clear understanding of outcomes and the statistical methods behind them. As such, it is vital that tools are identified that have been through a stringent validation process so that there is confidence in the data. Furthermore, the data that is produced must be amenable to a wide range of specialist and non-specialist stakeholders alike. These tools can be developed in-house, but this is often a prohibitively expensive undertaking. The TOPS form alone cost over £400,000 to develop and has demonstrated limited applicability to the substance misuse field as a whole. Put simply, the selection of a poor outcome tool will lead to poor commissioning, a lack of accountability in the service provider and poorer treatment response in clients seeking their help. This cannot be considerable acceptable when the responsibility of commissioners lies with quality assurances to the public purse and the vulnerable people who seek support from these vital services.