Comparison of indices of clinically meaningful change in Child and Adolescent Mental Health Services (CAMHS): difference scores, reliable change, crossing clinical thresholds and "added value" ; an exploration using parent rated scores on the SDQ.

Running head: Comparison of indices of clinically meaningful change in CAMHS

Miranda Wolpert, Evidence Based Practice Unit (EBPU), UCL and the Anna Freud Centre, 21 Maresfield Gardens, London, NW3 5SD

Anke Görzig, EBPU, UCL and the Anna Freud Centre, 21 Maresfield Gardens, London, NW3 5SD

Jessica Deighton, **Corresponding Author,** EBPU, UCL and the Anna Freud Centre, 21 Maresfield Gardens, London, NW3 5SD

Phone: 020 7443 2218

Email: Jessica.Deighton@annafreud.org

Andrew JB Fugard, EBPU, UCL and the Anna Freud Centre, 21 Maresfield Gardens, London, NW3 5SD

Robbie Newman, Child Outcomes Research Consortium (CORC), 21 Maresfield Gardens, London, NW3 5SD

Tamsin Ford, University of Exeter Medical School, Veysey Building, Salmon Pool Lane, Exeter EX2 4SG

Abstract

Background: Establishing what constitutes clinically significant change is important both for reviewing the function of services and for reflecting on individual clinical practice. A range of methods for assessing change exist, but it remains unclear which are best to use and under which circumstances.

Method: This paper reviews four indices of change (difference scores (DS), crossing clinical threshold (CCT), reliable change index (RCI) and added value scores (AVS) drawing on outcome data for 9,764 young people from child and adolescent mental health services across England.

Results: Looking at DS, the t test for time one - time two scores indicated a significant difference between baseline and follow up scores, with a standardised effect size of $d = 0.40$. AVS analysis resulted in a smaller effect size of 0.12. Analysis of those crossing the clinical threshold showed 21.2% of cases were classified as recovered, while 5.5% were classified as deteriorated. RCI identified 16.5% of cases to show reliable improvement and 2.3% of cases to show reliable deterioration. Across RCI and CCT 80.5% of the pairings were exact (i.e., identified in the same category using each method).

Conclusions: Findings indicate that the level of agreement across approaches is at least moderate; however, the estimated extent of change varied to some extent based on the index used. Each index may be appropriate for different contexts: CCT and RCI may be best suited to use for individual case review; whereas DS and AVS may be more appropriate for case- mix adjusted national reporting.

Keywords: Indices of change, child mental health, Difference Scores, Crossing Clinical Threshold, Reliable Change Index, Added Value Scores

Key Practitioner Message:

- There is a potential for different approaches to measuring change in symptoms over time to lead to different conclusions about the effectiveness of services: therefore it is inadvisable to make automatic summative judgements of services based on one approach or metric.
- It is important to use the most appropriate method corresponding to the specific question being asked. For example, CCT and RCI at the individual case level to review changes for individual children and families; standardised effect sizes (such as the AVS) to compare populations with a similar case mix and provide comparisons at the national policy level.
- Triangulation with other information – including service satisfaction, therapeutic alliance and functioning in everyday life – is an essential pre-requisite if such data are to be used in meaningful ways.
- It is essential that practitioners, service managers, policy makers, commissioners and service users themselves understand the differences in these metrics so that they can appropriately make sense of outcome data presented to them.

Introduction

Determining the optimal method to estimate the outcome of interventions in routine care, for both individual practice and service development, is an international policy priority (Department of Health, 2011; Dreyer, Tunis, Berger et al., 2010). A fundamental and largely unresolved issue is consensus about what constitutes a *clinically meaningful* change, with various proposed indices (Clark, Fairburn, & Wessely, 2008; Lambert & Ogles, 2009). Each currently available method has limitations. One common method is to look at basic *difference scores* (DS), between mean before and after scores on questionnaires. Typically, difference scores are formally tested using inferential statistics such as t-tests, which fail to adjust for attenuation, regression to the mean and random fluctuation. Effects observed, therefore, often represent an over-estimate of the level of change (Ford, Hutchings, Bywater et al., 2009) and are best understood in comparison with a matched group in research studies. Furthermore, statistical significance does not equate to clinical significance. An increase of one or two points on a questionnaire can be statistically significant but not noticeable or meaningful to the child or family. Conversely, families may report changes that have profoundly impacted on their lives, which may not be statistically significant, either because large changes may not reach statistical significance in a small sample and/or the measure used may not adequately detect the changes of importance to the family.

Converting the difference to a standardised effect size (e.g. Cohen, 1988, 1992) makes it more interpretable but the correspondence between this and clinical significance is also not entirely clear. An alternative, clinically intuitive, method is to measure *recovery*, defined as crossing a clinical threshold (CCT) on the measure used whereby the individual moves from the category of having a clinically significant problem to not having one. Patients who move beyond a clinical threshold (from present to absent) are considered recovered. However, someone with a high baseline score may have a dramatically improved score at follow-up,

but not cross the clinical cut point, while a much smaller change in someone whose initial score was close to the cut point might produce *recovery*, as defined above. In addition, the choice of cut-off points is always subjective, balancing a need to be sensitive to features of a condition indicating clinical levels of need, whilst also being specific enough to prevent "false alarms" where people are without a need for clinical input. Moreover, for some conditions such as autism spectrum conditions, for example,, where core difficulties are unlikely to change significantly, what constitutes *recovery* is not entirely clear or cannot easily be quantified on general outcome measures.

Another method for establishing whether significant change has occurred is the *reliable change index* (RCI; Christensen & Mendoza, 1986; Jacobson, Follette and Revenstorf, 1984; Jacobson & Truax, 1991; McNemar, 1960). This estimates the amount of change required in an outcome measure to confidently conclude that the change observed is not solely attributable to measurement error. It has been proposed as a way to meaningfully classify individuals as *improved, unchanged* or *deteriorated*. However, changes in scores that are meaningful to the individual do not always meet the criteria for reliable change and changes that are not clinically meaningful may be identified as "improved"t. Furthermore, this approach does not account for attenuation due to use of repeated measures, regression to the mean or spontaneous improvement (Hageman & Arrindell, 1993).

A recent attempt to address these issues is the *added value score* (AVS) for the Strengths and Difficulties Questionnaire (SDQ; Ford et al., 2009.)(.) This computerised algorithm uses comparative epidemiological data from young people with research diagnoses of psychiatric disorder and/or parental help-seeking from primary care or schools to develop a linear equation that estimates an added value score for targeted and clinical child mental health interventions at group level. This approach mirrors that used to assess and compare the effectiveness of schools, which has also been recognised to have the following

limiatations (Leckie & Goldstein, H 2011). .. The AVS accurately predicted the effect size obtained by control and interventions groups on the SDQ in two randomised controlled trials (Ford et al., 2009; Rotheray et al., 2014). However, it is not yet clear if it works for all groups of children or interventions; both trials used to test parenttraining with young children at high risk of conduct disorder. Moreover, the AVS is not appropriate for use with individual cases or very small numbers of cases (Ford, et al., 2009).

The determination of meaningful change for children and families attending CAMHS may be further complicated by the need to consider multiple perspectives and the impact of development on suitable clinical cut-offs. Whilst full discussion of the implicatios of these aspects are beyond the scope of this paper, but are covered in more depth elsewhere (Wolpert et al 2014;) (http://www.cypiapt.org/routine-outcome-monitoring/routine-monitoring-outcome.php)

The current paper reviews these different methods of estimating clinical change outlined above (i.e., DS, CCT, RCI, and AVS) in order to advise those dealing with routine-collected outcome data of the strengths and limitations of each and to suggest possible ways forward. Practical illustrations are provided of the different results each method achieves based on a large dataset of routinely-collected outcome data from child and adolescent mental health services (CAMHS) in England and Scotland. These data were collected by the Child Outcomes Research Consortium (CORC; Wolpert, Ford, Trustam et al., 2012).

Method

Participants and Procedure

The current sample was drawn from a wider dataset of demographic details and mental health outcome information collected locally by CAMHS that were part of CORC

(Wolpert et al., 2012). Since this is anonymised routinely-collected outcome data, no research ethical approval was required.

The data used in the current analyses is comprised of 9,764 young people from 58 services across England, whose parents completed the SDQ at intake (time one) and 4-8 months later (time two) as part of their routine treatment. Of these young people, 97.5% reported their age at referral, which ranged from 0 to 18 years (*M*=10.81, *SD*=3.43), 44.8% were female (gender reported by 98.5%), while 74.7% of the 69.6% who reported their ethnicity were white British. Based on available data (*n*=3,108), emotional problems were the most common specified presenting problem (57.9%), followed by conduct problems (20.2%), 'other' presenting problems (16.8%), hyperkinetic disorder (11.1%), autistic spectrum disorder (9.8%), deliberate self-harm (6.0%), eating disorder (5.6%), learning disability (3.9%), habit disorder (3.7%), developmental disability (3.4%), psychotic disorder (1.1%) and substance use disorder (1.0%). The most common type of professional grouping (available data n=3,283) involved was clinical psychology (38.9%), followed by medical (23.0%), nursing (10.7%), primary mental health worker (13.9%), psychotherapy (8.7%), family therapy (7.5%), 'other' professional input (7.4%), social work (6.7%), occupational therapy (3.2%), creative therapy (2.3%), educational other (2.1%), counselling (1.5%) and education psychology (0.8%).

The most common intervention (available data n=3,273) was 'other therapeutic intervention' (31.1%), followed by counselling (28.3%), cognitive behavioural therapy (17.0%), parent training (8.0%), family therapy (7.4%), drug treatment (5.6%), child psychotherapy (2.7%), 'other' parent intervention (2.4%), creative therapy (3.6%) and neuropsychological intervention (1.2%).

Measure

The SDQ (SDQ; Goodman, 1997 and 2001) is a widely-used and well-validated measure composed of four subscales that assess distress (conduct problems, emotional problems, peer problems and hyperactivity) and one subscale assessing strengths (pro-social skills). Each subscale is composed of five items which are rated on a scale from 0 (*not true*) to 2 (*certainly true*). The total difficulties score is computed as a sum of the four difficulty subscales, while a separate impact score is calculated from the impact supplement where the informant rates child distress and the impact of difficulties on home life, friendships, classroom learning, and leisure activities are combined to form the impact score. General population data (www.sdqinfo.org) demonstrates small decreases in scores with age amounting to less than 0.5 points in total difficulties score between primary and secondary school age children, and thus do not suggest the need for differentiated clinical cut points.

Analysis

The following four analyses of change for DS, CCT, RCI and AVS were undertaken. In order to compare these approaches, all analyses were carried out using the total difficulties scales of the parent-reported SDQ. The only exception to this was the AVS, which is calculated using all scales from the total difficulties scale and the impact score.

Difference scores (DS). DS were based on the mean change in the sample scores between time one and time two; this change was tested using a paired-measured t-test and converted into standardised effects size (Cohen's d; Cohen, 1988, 1992).

Crossing clinical thresholds (CCT). The clinical threshold for identifying the presence of a psychiatric disorder through the parent rated SDQ total difficulties score of ≥ 17 was used in the current analysis (Goodman, 1997, 2001; Goodman, Ford, Simmons et al., 2000; Improving Access to Psychotherapy, 2012). CCT applies to individual cases and is then reported on a group level as the proportion that crosses the threshold.

Reliable change index (RCI). RCI, as proposed by Jacobson and Truax (1991), is a function of the reliability of the scale (usually Cronbach alpha (used here) or test-retest reliability (see Table 3). Like CCT, RCIs are computed for individual cases, but are reported for groups as the proportion having reliably changed.

Added value scores. The algorithm for calculating added value scores was developed for the parent version of the SDQ on a sample of children with a similar profile to those who attend CAMHS but the majority of whom did not receive treatment (Ford, et al., 2009). The algorithm was developed empirically to determine the optimal SDQ baseline variables with which to predict the total difficulties score at 4-8 months follow up, aiming to remove the influence of random fluctuation, regression to the mean and spontaneous improvement (see http://www.sdqinfo.com/c5.html and Table 3). The AVS is designed to be reported at the group rather than individual level. It is typically reported as a mean standardised effect size in SD units. A mean added value score of zero indicates that the population shows no change over that expected in an untreated sample, a negative score indicates that the change in scores is worse than predicted, and a positive score suggests the change in scores is better than predicted, i.e. they are reporting fewer difficulties than would be expected if left untreated.

Analysis was carried out using SPSS 21. A McNemar test was employed to see if CCT suggested a statistically significant difference in the proportions with scores above the clinical threshold at time one and time two and Kappa test for chance-corrected agreement was used to examine the concordance between young people allocated to “improved”, “no change” or “deteriorated” categories according to CCT and the RCI.

Results

Difference Score

The mean difference between parent rated SDQ total difficulties score at time one (*M* = 18.96, *SD* = 7.14) and time two (*M* = 16.10, *SD* = 7.98) was 2.86 SDQ points, which

constituted a statistically significant difference (t (9,763) = 46.12, $p < .001$). This corresponds with a standardised effect size of $d = 0.40$ (95% *CI* 0.26 to 0.54) (Cohen, 1992), calculated using the standard deviation at baseline (Becker, 1988.)

Crossing Clinical Threshold

Using the proposed cut-off point of a total difficulties score ≥ 17, 21.2% of cases were classified as recovered (i.e., have crossed the threshold from clinical to non-clinical) while 5.5% have crossed the clinical threshold from non-clinical to clinical, and, therefore, would be considered to have deteriorated (see Table 1). Readers may be surprised that in a sample of children attending mental health services "only" 63% scored above the clinical cut point of 17. This is a well-known finding and is likely to be partly due to measurement error inherent in the questionnaire and partly because the SDQ focuses on general emotions and behaviour, and does not have specific questions on some of the conditions that are commonly seen in mental health services such as Autism Spectrum Conditions and Eating Disorders, although children with these conditions tend to score substantially higher than the population mean due to co-morbidities.

INSERT TABLE 1 ABOUT HERE

Reliable Change

Given the standard deviation of the sample at the first measurement point (*SD* = 7.14) and the reliability of the parent rated SDQ from a nationwide epidemiological sample of 10,438 British 5-15 year olds (Cronbach's α = 0.82; Goodman, 2001), 16.5% (1,610 cases) were found to show an improvement that was statistically significant at the 5% significance level(i.e., showed a decrease in SDQ total difficulty scores of 8.36 or more) while 2.3% (224 cases) had deteriorated (i.e., showed an increase in SDQ total difficulty scores of 8.36 or more) between time one and time two.

Agreement between classification of cases as showing 'improvement', 'deterioration' or 'no change' based on RCI and CCT was moderate (Kappa = .472; p<.001), and in fact 80.5% of the pairings were exact and there were no cases that "improved" according to one system yet "deteriorated" according to the other (see Table 2). Therefore, all disagreements were between no change and deterioration/improvement. 7.88% of those identified as showing no change on CCT are identified as improved or deteriorated on RCI, whereas 16.86 % of those showing no change based on RCI are identified as improved or deteriorated based on CCT. These cases were termed 'discordant cases'. Discordant cases that were identified as having changed based on RCI but not based on CCT (n = 564) had a mean SDQ Total Difficulties score at time one of 20.87 (SD = 8.27) and a mean score at time two of 13.88 (SD = 10.43). These cases changed by as much as 19 raw score points (range in change: -16 to 18; mean change = -6.98, SD = 8.34), and most (n = 466) showed changes in terms of reduction in symptom scores from time one to time two. Cases identified as having changed based on CCT but not on RCI (n = 1,337) had a mean time one Total Difficulties Score of 17.67 (SD = 2.93) and a mean time two Total Difficulties Score of 15.34 (SD = 3.08). Over half of the cases scored within 2 points of the clinical threshold at time one (score range 15 ≥ 19: n = 788). The cases changed by a maximum of 8 raw score points from time one to time two (range in change: -8 to 8; mean change = -2.34, SD = 5.11). Again, the majority of cases (n= 925) showed a reduction in symptom scores over time.

INSERT TABLE 2 ABOUT HERE

Added Value Score

The AVS for the parent-rated SDQ has a mean standardised effect size of 0.12 (95% *CI* .096 to .14, *SD* = 1.15), which suggests that service users on average show modest improvements over the general population control group of young people and children. There

is a substantial reduction in effect size calculated using simple change scores (0.40) to that obtained using the AVS (0.12), which may indicate the extent to which random fluctuation, regression to the mean and attenuation inflate change scores or that the AVS may underestimate change.

Discussion

In this paper we have reviewed four different approaches to assessing the outcome of interventions in routine care (DS, CCT, RCI and AVS) on the same sample and using the same measure to illustrate the heterogeneity in reported outcomes based on different methods. All indices suggested that the majority of young people experienced improvements in their mental health as reported by parents between the two data collection points. The extent of the change, and in relation to the RCI and CCT, who was classified as deteriorated or improved versus 'no change', did vary based on the approach taken. However, there was a moderate agreement between the different approaches and discordant cases were not between recovered and deteriorated but between either of these and no change.

DS converted to a standardised effect size provide the simplest and perhaps the most intuitive means of assessing group level changes in scores on outcome measures before and after treatment, but may be overly simplistic. In particular, statistical tests are sensitive to sample size with small samples entailing less confidence in results, whilst with large samples, even very small pre-post difference can yield statistically significant results that do not equate to clinically meaningful change. Reporting DS in conjunction with a standardised effect size, such as Cohen's d (Cohen, 1992), can ameliorate this but does not take account of regression to the mean, attenuation and spontaneous improvement, as clearly illustrated by the difference in effect sizes calculated according to the DS (.40) and AVS (.12). Similar differences were obtained when the AVS and DS were tested in the intervention and control arms of two randomised controlled trials (Ford et al., 2009; Rotherway et al., 2014). DS do

not inherently incorporate comparative information in the way AVS does, by adjusting the estimated size of change using information from a proxy control group. Therefore, the DS approach should only be used when a control group, or a suitable comparison group, is available.

CCT can be applied at the level of the individual (unlike DS and AVS) and makes a distinction between scores reflecting a clinical, versus a functional, population, which arguably reflects change with the most clinical relevance or practical significance. However, it does not make a distinction between smaller and larger changes. Random fluctuations in scores closely clustered around the chosen cut off will be reported as 'significant,' whilst large changes (positive and negative) which fail to cross the threshold will go unreported. In addition, it does not detect even large changes in cases where no threshold is crossed. For many types of childhood psychopathology, there are no clear, clinically indicated thresholds, while for some measures such as the Conners scales, cut points differ by gender and age, which can lead to errors. Furthermore, this approach equates a change in symptomatology with good clinical outcome. Disorders such as autism spectrum conditions may gain substantial improvements in quality of life and functioning, but would not be expected to necessarily reduce core psychopathology. CCT should, therefore, be used in combination with RCI: this combination could usefully inform individual case review.

Unlike CCT, RCI takes account of change anywhere in the scoring range. Moreover, it gives an indication of individual level change whilst removing an element of measurement error and is not tied to a particular measure, control group or previously established norms. Despite these strengths, RCI has a number of issues. For example, it is not sensitive to small changes that may be clinically meaningful to young people and families (Wise, 2003, 2004). Cases showing a reduction of up to eight raw score points from above the clinical threshold at time one to within the 'normal' range at time two on the SDQ Total Difficulties scale would

not be identified based on RCI. Furthermore, RCI does not take into account regression to the mean and so may function less effectively with those scoring at the extremes (i.e very high or very low) at baseline (Hageman & Arrindell, 1993). RCI could provide useful client-level information alongside CCT. However, to assess service level change (i.e., percent having reliably changed), because of the lack of account for regression to the mean, we recommend that RCI is considered in comparison to a control group, similar clinical services or the AVS.

Because the AVS makes an implicit comparison to a population not receiving treatment it does not necessarily require comparison to other services to aid interpretation of service-level change. A cut point of 0.15 has been suggested in relation to the AVS with recent policy precedents but there is no empirical evidence to date to support this. Moreover, the AVS is calculated based on the assumption that an epidemiological "high risk" sample is an appropriate control group. A test of the AVS in two trials, both of parenting interventions in early childhood suggest that the AVS functions well in this group (Ford et al., 2009; Rotheray, Racey, Rodgers et al., 2014), but the score has yet to be tested with other interventions or in other age groups. The severity of psychopathology seems to be a consistent predictor of attendance at mental health services (Angold, Erkanli, Farmer et al., 2002; Ford, Vostanis, Meltzer et al., 2007) while both tests of the AVS involved high-risk samples rather than CAMHS attenders.

To our knowledge, this is the first demonstration using clinical data to directly compare different methods of estimating clinical change and it benefits from a large sample size. However, some limitations should be acknowledged. As the analysis was based on routinely-collected data where information about the sampling frame was missing, we cannot know how representative the sample is of the children that it was drawn from.

We make no claims about the change rates themselves on this basis, but it may be that this also may have led to a biased sample in terms of comparing approaches. It is not known how representative the data from the current sample is, although it is consistent with CAMHS mapping in terms of proportion of presenting problems reported (www.childrensmapping.org.uk) and generalizability is arguably not key to these illustrative analyses. In addition, we only considered one measure for illustrative purposes – the SDQ – and only used the four most common forms of change indices; further work in this area might usefully consider other measures and additional change indices, although the AVS is currently only available for the parent-reported SDQ. In particular, development of added value scores for other outcomes measures would be welcomed. Further research comparing changes in such measures against external clinical measures of change would also be useful.

Conclusion

If we treated the sample used in this review as a service, it is clear that there is the potential for different approaches leading to different conclusions about the effectiveness of the service. The CCT or RCI indicate improvement in approximately 20% of the sample, although they will, in a small proportion of instances, identify different cases as improved. The starkest contrast is in the standardised effects sizes for DS and AVS. The warning from this work is particularly apposite given the recent difficulties around heart surgery in Leeds (The Lancet, 2013). It is clearly unwise to make automatic summative judgements of services based on one approach or metric: different approaches give different answers. It is important to use the most appropriate method corresponding to the specific question being asked (e.g., service effectiveness versus individual client improvement). It is already clearly demonstrated in the literature that difference scores should not be judged by themselves but in conjunction with measures assessing outcomes that are of concern to the patient, such as

service satisfaction, therapeutic alliance and functioning in everyday life (Kazdin, 1999; Lunnen & Ogles, 1998; Wise, 2004).

We would fully endorse this approach and go further to say that triangulation with other data is an essential pre-requisite for outcome data to be used in meaningful ways. However, further empirical investigation into the best combination of approaches is required. Future studies might consider use of outcome measures and satisfaction measures, or possibly how to incorporate qualitative information about services, such as that developed by the Quality Network for Inpatient CAMHS (QNIC; CAMHS, 2013).

INSERT TABLE 3 ABOUT HERE

Table 3 summarises how the indices studied might be most meaningfully applied. We suggest that CCT and RCI may be most useful at the individual case level to review changes for individual children and families in the light of other information. Whereas standardised effect sizes, when used to compare populations with similar levels of initial severity, and AVS, when used to compare case mix adjusted services, may be more appropriate at the national policy level. Clinical outcomes at an individual level should prompt practitioners to reflect on the nuances of a case, including scrutiny of responses to individual items as well as overall indices of change. If RCI and CCT are employed and give different answers, the practitioner should consider why; this could be a focus for discussion about progress with the family.

Even this represents a far from perfect solution. At the very least it is essential that service managers, policy makers, commissioners and service users themselves understand the differences in these metrics so that they can appropriately make sense of outcome data presented to them. Similar studies of the impact of case-complexity factors on the other

indices would be informative and should be a focus of further research.

Funding

Work on developing this paper was funded by the Department of Health Policy Research Programme. The views expressed are not necessarily those of the Department. Authors have no other interests (financial or otherwise) relevant to the submitted work.

Declaration of interest

Miranda Wolpert is paid director and Tamsin Ford is an unpaid director for the Child Outcomes Research Consortium.

Author contributions

Miranda Wolpert provided the specific conception of the paper and gave final approval of the article to be published. Anke Görzig developed the design, carried out the majority of the analysis for the article and developed the initial draft of the paper. Jessica Deighton oversaw the analysis, provided substantial revision to the paper and contributed to the interpretation of the data. Andrew Fugard provided critical comments and revision, particularly for the analysis included in the article. Robbie Newman carried out additional data analysis and summary statistics for description of the sample. Tamsin Ford provided critical comments and substantive revisions the plans for the paper, and to subsequents drafts of the manuscript..

Acknowledgements

The authors would like to thank all members of the Child Outcomes Research Consortium; the CORC committee at the time of writing (includes M.W. and T.F.): Ashley Wyatt, Alison Towndrow, Duncan Law, Evette Girgis, Julie Elliott, Ann York, Mick Atkinson and Alan Ovenden. The CORC central team at the time of writing

(includes M.W.): Isobel Fleming, Jenna Bradley, Rachel Argent, Robbie Newman, Slavi Savic and Thomas Booker.

The authors would also like to thank members of the Policy Research Unit in the Health of Children, Young People and Families: Terence Stephenson, Catherine Law, Amanda Edwards, Ruth Gilbert, Steve Morris, Helen Roberts, Russell Viner, and Cathy Street. This is an independent report commissioned and funded by the Policy Research Programme in the Department of Health. The views expressed are not necessarily those of the Department.

References

Angold, A., Erkanli, A., Farmer, E. M., Fairbank, J. A., Burns, B. J., Keeler, G., & Costello, E. J. (2002). Psychiatric disorder, impairment, and service use in rural African American and white youth. *Archives of General Psychiatry, 59*(10), 893-901. doi: yoa10103 [pii]

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41,* 257-278.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler.* (5th ed.) Berlin: Springer.

CAMHS, Q. I. N. i. (2013). Retrieved March 20th, 2013, from http://www.rcpsych.ac.uk/workinpsychiatry/qualityimprovement/qualityandaccreditation/childandadolescent/inpatientcamhsqnic.aspx.

Christensen, L. & Mendoza, J. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behaviour Therapy, 17* (3), 305-308.

Clark, D. M., Fairburn, C. G., & Wessely, S. (2008). Psychological treatment outcomes in routine NHS services: A commentary in Stiles et al (2007). *Psychological Medicine, 38*(5), 629-634.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Department of Health. (2011). *NHS Outcomes Framework 2012-13*. London: Department of Health.

Dreyer, N. A., Tunis, S. R., Berger, M., Ollendorf, D., Mattox, P., & Gliklich, R. (2010). Why observational studies should be among the tools used in comparative effectiveness research. *Health Affairs, 29*(10), 1818-1825. doi: 10.1377/hlthaff.2010.066629/10/1818 [pii]

Ford, T., Hutchings, J., Bywater, T., Goodman, A., & Goodman, R. (2009). Strengths and Difficulties Questionnaire Added Value Scores: evaluating effectiveness in child mental health interventions. . *British Journal of Psychiatry, 194*(6), 552-558.

Ford, T., Vostanis, P., Meltzer, H., & Goodman, R. (2007). Psychiatric disorder among British children looked after by local authorities: comparison with children living in private households. *British Journal of Psychiatry, 190,* 319-325. doi: 190/4/319 [pii] 10.1192/bjp.bp.106.025023

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 38*(5), 581-586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(11), 1337-1345. doi: S0890-8567(09)60543-8 [pii]; 10.1097/00004583-200111000-00015

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry, 177*, 534-539.

Hageman, W. L., & Arrindell, W. A. (1993). A further refinement of the reliable change index by improving the pre-post-difference score: Introducing the RCID. *Behaviour Research and Therapy, 51*, 693-700.

Improving Access to Psychotherapy (2012). Routine Outcome monitoring as part of CYP IAPT. Retrieved from http://www.iapt.nhs.uk/cyp-iapt/routine-outcome-monitoring-as-part-of-iapt/

Jacobson, N.S., Follette, W.C. & Revenstorf, D. (1984). Psycotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behaviour Therapy, 15* (4), 336-352.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12-19.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*(3), 332-339.

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research, 19*(4-5), 493-501. doi: 10.1080/10503300902849483

Leckie, G., & Goldstein, H. (2011). A note on 'The limitations of school league tables to inform school choice'. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174(3), 833-836.

Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*(2), 400-410.

Rotheray, S., Racey, D., Rodgers, L., Berry, V., McGillway, S., & Ford, T. (2014). *The effectiveness of the SDQ Added Value Score*. *Child and Adolescent Mental Health*. doi: 10.1111/camh.12059

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *Canadian Journal of Psychiatry, 52*(2), 121-128.

Swets, J. A. (1973). The Relative Operating Characteristic in Psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science, 182*(4116), 990-1000. doi: 182/4116/990 [pii] 10.1126/science.182.4116.990

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychol Bull, 99*(2), 181-198.

The Lancet. (2013). Children's heart surgery in Leeds, UK. *Lancet, 381*(9874), 1248. doi: 10.1016/S0140-6736(13)60824-7S0140-6736(13)60824-7 [pii]

Wise, E. A. (2003). Psychotherapy outcome and satisfaction methods applied to intensive outpatient programming in a private practice setting. . *Psychotherapy: Theory, Research, Practice, Training, 40*, 203-214.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: a review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82*(1), 50-59. doi: 10.1207/s15327752jpa8201_10

Wolpert, M., Ford, T., Trustam, E., Law, D., Deighton, J., Flannery, H., & Fugard, A. (2012). Patient-reported outcomes in child and adolescent mental health services (CAMHS): use of idiographic and standardized measures. *Journal of Mental Health, 21*(2), 165-173. doi: 10.3109/09638237.2012.664304

Wolpert, M., Deighton, J., De Francesco, D., Martin, P., Fonagy, P., & Ford, T. (2014). From 'reckless' to 'mindful'in the use of outcome data to inform service-level performance management: perspectives from child mental health. BMJ quality & safety, bmjqs-2013.

Tables

Table 1. Percentages of clinical cases at time one and time two according to threshold criteria.

		Clinical, time two		
		No	Yes	*Total*
Clinical, time one	No	31.9%	**5.5%**	37.4%
	Yes	**21.2%**	41.4%	62.6%
Total		53.1%	46.9%	100.0%

Notes: N = 9,764, McNemar χ^2 = 897.92 , p < .001.

Table 2. Agreement of RCI and CCT classifications.

		CCT			
		Negative change (deterioration)	No change	Positive change (improvement)	Total
RCI	Negative change (deterioration)	126	98[a]	0	224
	No change	412[b]	6.593	925[b]	7,930
	Positive change (improvement)	0	466[a]	1,144	1,610
Total		538	7,157	2,069	9,764

[a] = change identified by RCI but not by CCT; [b] = change identified by CCT but not RCI

Table 3: A summary of the four approaches

Method	Strengths	Weaknesses	Statistics or methods used	Recommended use
Difference Scores	Simple to understand and calculate, can be converted into a standardised effect size.	Fails to account for regression to the mean, attenuation and random fluctuation (when no control group is present). Does not necessarily indicate clinical significance.	Effect size $d = (M_{t1}-M_{t2})$ (Bortz, 1999; Cohen, 1988; Becker; 1988.)	For national reporting and consideration against published data where same level of severity is present at outset or for comparison with another similar group.
Crossing Clinical Threshold (CCT)	Attempts to determine a score that best distinguishes between a clinical and a functional population.	Does not differentiate between smaller and larger changes as long as threshold is crossed. Clinical cut point may be difficult to determine for some disorders.	Triangulation with diagnostic criteria and using Receiver Operating Characteristics (ROC) procedures such as sensitivity, specificity (Streiner & Cairney, 2007; Swets, 1973, 1986) and cross-validation with other instruments.	Combined with RCI for individual case review along with other triangulated data.

Reliable Change Index (RCI)	Attempts to measure statistically reliable change in the absence of measurement error.	Low sensitivity to small but clinically meaningful changes. Does not necessarily indicate clinical significance.	RCI= M_{t1}-M_{t2}/SE_{diff} with SE_{diff} = SQRT(2xSE^2) with SE = SD x SQRT(1 – rel) $\Rightarrow$ SE_{diff} = SQRT(2 x (SD x SQRT(1 – rel))^2). Significant change at the 95% confidence level: (RCI $\geq$ 1.96)	Combined with CCT for individual case review along with other triangulated data.
Added Value Score (AVS)	Attempts to take account of regression to the mean and spontaneous improvement to determine what amount of change has taken place that is due to an intervention as opposed to change as an artefact or other individual or contextual factors.	Is bound to the population and measure for which the algorithm was developed Does not necessarily indicate clinical significance.	Raw SDQ AVS (in SDQ points) =2.3 + 0.8 x T1 total difficulties score + 0.2 x T1 impact score - 0.3 x T1 emotional difficulties subscale score - T2 total difficulties score Divide by standard deviation to obtain an effect size.	For benchmarking across services with similar case mix.

Notes: M = Mean; n = number in the sample at one measurement point; SD = Standard Deviation; SQRT = Square Root; T1 = first measurement point; T2 = second measurement point