Evaluation of the Strengths and Difficulties Questionnaire Added Value Score as a method for estimating effectiveness in child mental health interventions

Tamsin Ford[*1], Judy Hutchings[2a], Tracey Bywater[2b], Anna Goodman[3], Robert Goodman[4]

1. Clinical senior lecturer, Institute of Health Services Research, Peninsula Medical School, St Luke's Campus, Heavitree Road, Exeter EX2 8UT

2 a. Professor & 2b, Project trial coordinator, School of Psychology, Bangor University, Brigantia Building, College Road, Bangor, Gwynedd, LL57 2DG, Wales

3. PhD student, Department of Epidemiology and Public Health, London School or Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

4. Professor of Brain and Behaviour Medicine, Department of Child and Adolescent Psychiatry, Kings College London: Institute of Psychiatry, De Crespigny Park, London SE5 8AF.

*corresponding author: tamsin.ford@pms.ac.uk

Word count 4660

Summary

Background

Routine outcome monitoring may improve clinical services but remains controversial, partly because the absence of a control group makes interpretation difficult.

Methods

We developed an "Added Value" score using epidemiological data on the Strengths and Difficulties Questionnaire. We tested whether it correctly predicted the effect size for the control and intervention groups in a randomized controlled trial.

Findings

As compared to *a priori* expectation of zero, the Added Value Score applied to the control group predicted an effect size of -0.03 (-0.30 to 0.24; t=-0.2, p=0.8). As compared to the trial estimate of 0.37, the Added Value Score applied to the intervention group predicted an effect size of 0.36 (0.12 to 0.60; t=-0.1, p=0.9).

Interpretation

Our findings provide preliminary support for the validity of this approach as one tool in the evaluation of interventions with groups of children who have, or are at high risk of developing, significant psychopathology.

Introduction

Although it is clear that a variety of child mental health treatments are *efficacious* (i.e. have an impact under ideal trial conditions), there is still considerable doubt about the *effectiveness* of interventions for children with mental health problems in everyday practice.[1,2] Given the recent expansion of mental health services for children in Great Britain, this uncertainty should preoccupy those involved in service delivery, development and policy.[2] The publication of routinely collected data on post-operative mortality in cardiac surgery may have contributed to a reduction in post-operative mortality; although routine outcome monitoring is not without controversy in this and other specialties.[3] Despite the misgivings of some mental health practitioners, routine outcome monitoring has been recommended as a way of driving up the standards of Child and Adolescent Mental Health Services (CAMHS).[4] The lack of a control group for routinely collected outcome data means that any change after treatment cannot be directly attributed to the intervention provided, as other factors may also have changed in the interim period. As most CAMHS attenders will score at the higher end of psychopathology scales, we would expect their psychopathology scores to reduce in the short term because of regression to the mean, attenuation and the fluctuating nature of most childhood psychopathology. Regression to the mean occurs due to random measurement error, so that the second measurement of low and high scorers on any scale will tend to score nearer the mean.[5] Attenuation refers to the tendency identified in epidemiological studies for people to report more problems in the first than subsequent interviews, perhaps because of respondent fatigue.[6] Childhood psychiatric disorders have a chronic and fluctuating course, and as people are often referred when their problems are at a peak, in the short term the severity of a child's difficulties are likely to reduce with or without active intervention, despite substantial long term continuity in most types of difficulties.[7] Could epidemiological data about the longitudinal course of childhood psychopathology in the community be used to predict expected change in much the same way that growth charts are currently used for height, weight and body mass index?[8,9] Adjusting for expected change would allow services to calculate a more realistic estimate of the "added value" of their interventions. We used data from a longitudinal study of childhood psychopathology in the community[10] to develop a computer algorithm that we

then tested against data from a randomized controlled trial.[11] If the computer algorithm worked as a measure of added value, then it should be able to correctly predict the outcomes of the intervention and control groups in that trial. If we could demonstrate that the algorithm worked as predicted on data from randomized controlled trials, then it would support the case for using the same algorithm to assess intervention-related change in clinical practice.

Method

Development of the SDQ Added Value Score

The Added Value Score was derived from scores on the Strengths and Difficulties Questionnaire (SDQ) completed by parents of children aged 5-16 years participating in the British Child and Adolescent Mental Health Survey 2004 (n=7977), and the follow up 4-8 months later.[10] The follow up study involved all those who were assessed as having a psychiatric disorder at baseline (n=705) and a random sample of those without (n=926). Nearly all (96%) parents participating in the baseline survey agreed to be contacted again, and the response rate for the follow up survey was 72%.

The SDQ is a well-validated 25-item screening questionnaire composed of five scales that assess behaviour problems, hyperactivity, emotional symptoms, peer problems and pro-social skills.[12] Responses to questions from the first four subscales are added to give a total difficulties score. Ratings of child distress and the impact of difficulties on home life, friendships, classroom learning, and leisure activities are combined to form the impact score. The follow up version of the SDQ (see www.sdqinfo.com) asks whether any difficulties the child had at baseline have changed, using a five-point Likert type scale (much worse, a bit worse, about the same, a bit better, much better). Questions forming the basis of the total difficulties and impact scores were identical at both time points, except that the baseline questionnaire asked about difficulties within the previous six months while the follow up questionnaire was restricted to the previous month.

Parents and young people aged 11 years or over participating in the British Child and Adolescent Mental Health Survey 2004 also completed the Development and Well-Being

Assessment in the baseline survey (DAWBA).[13] The DAWBA is a structured diagnostic interview that was administered by lay interviewers. If the family agreed, a shortened version was mailed to the child's teacher. All informants were asked to describe any problem areas in their own words, using a series of prompts, and a small team of experienced child psychiatrists used information from the structured questions and verbatim transcripts from all informants to allocate diagnoses of psychiatric disorder using ICD 10.[14] In the validation study of the DAWBA, there was excellent discrimination between community and clinical samples.[13] Within the community sample, children with DAWBA diagnoses differed markedly from those without a disorder in external characteristics and prognosis, while there were high levels of agreement between the DAWBA and case notes among the clinical sample (Kendall's tau b = 0.47-0.70).

When constructing the SDQ Added Value Score, we selected children from the follow up of the British Child and Adolescent Mental Health Survey 2004 who were either rated as having a psychiatric disorder (n=455) in the baseline survey or whose parents had contacted primary health care or teachers about mental health concerns within the previous year (n=437); given the substantial overlap between these groups, this identified a group of 609 children. We had chosen these selection criteria to identify a group as similar as possible to children who attend CAMHS. Follow up SDQ scores were influenced by the presence of a psychiatric disorder at baseline (+1.2 SDQ points, $p<0.001$) and contact with contact with primary health or teachers (+1.3 SDQ points, $p<0.001$), but not gender (more boys than girls attend CAMHS).

Some of these children (n=100, 16%) reported attendance at CAMHS during the follow up period, but given that their SDQ scores at the first attendance of CAMHS were not available, we were ignorant as to their position on the intervention trajectory. For example, a child with a score of 18 in the baseline survey, might then deteriorate acutely two months later to 24, prompting referral to CAMHS, but given preliminary intervention their score might improve to 20 by the 6 month research follow-up. This would lead to the child being 2 points worse at follow up even though there had been improvement

following preliminary intervention by CAMHS. The mean SDQ Added Value Scores of CAMHS attenders were significantly worse than those of children who reported no mental health contact (-2.0, standard deviation 5.1, versus +0.3, SD 4.6, $p<0.001$). Thus, we included CAMHS attenders the sample as their exclusion might have left a sample of children with milder difficulties who were less representatives of children requiring mental health services.

The computer algorithm was developed empirically by applying linear regression to the baseline SDQ scores of the 609 children to predict their follow up SDQ total difficulties scores as accurately as possible from their initial SDQ scores. We found that the independent predictors of follow up total difficulties score, using stepwise multiple regression, were the baseline scores for total difficulties, impact and emotional symptoms (more details available from the authors on request and on www.sdq.info.com). The SDQ Added Value Score is essentially the difference between the expected and observed outcome at follow up and is normally distributed, with a mean of zero and a standard deviation of 5 SDQ points. Scores greater than zero reflect better than predicted adjustment, while scores less than zero indicate worse than predicted adjustment. Added Value Scores showed a modest correlation with parents' reports of the change in their children's difficulties since the baseline survey (Spearmans's rho =0.30, $p<0.001$), but as Figure 1 illustrates the relationship between the two measures of change was broadly linear and in the expected direction.

Insert Figure 1 about here

We used stepwise linear regression to examine the extent to which "case mix" variables or context predicted the SDQ Added Value Score. Only 0.6% of the variance of the SDQ Added Value Score was accounted for by the wide range of "complexity" characteristics measured in the baseline survey, namely type and severity of diagnosis, age, gender, intelligence, physical health, maternal educational level, maternal anxiety or depression, family type, family function, family size, income, housing tenure and neighbourhood characteristics. In contrast, the variance in SDQ total difficulties explained by these same

characteristics was 35.9% at baseline and 24.2% at follow up, demonstrating that the influence of case complexity on the SDQ Added Value Score was very small in this sample, and is certainly much reduced compared to the influence of these characteristics on raw scores. This suggests that providing the SDQ Added Value Score is used with children who have or are at high risk for impairing psychopathology (because this mirrors the children that it was derived from), the function of the algorithm may not vary a great deal in different contexts.

Study design and participants of the Welsh Sure Start randomised controlled trial

This trial was selected to test the SDQ Added Value Score because it used the SDQ with the impact supplement, had a follow up 4-8 months later and detected a difference between the control and intervention groups. It was the only trial meeting all these criteria that we were able to locate by searching trial registries for trials using the SDQ as an outcome measure and by contacting researchers running trials of child mental health interventions. The trial tested the Incredible Years Basic Parenting Programme; a 12-week group intervention aimed at reducing behavioural problems in children.[14] Parents were randomly allocated on a 2:1 ratio to immediate or delayed treatment.[11] The program has a strong evidence-base in the prevention and treatment of conduct disorder, and is one of two treatments for conduct disorder specifically recommended by the National Institute for Health and Clinical Excellence.[16] The trial took place in eleven Sure Start areas in North and Mid Wales, delivering a standardised behavioural programme in community settings using their existing staff.[11]

The children were aged three and four years old and at risk of conduct disorder defined as scoring above cut off on one or both of the intensity or total problem scales on the Eyberg Child Behaviour Inventory (ECBI).[17] The trial reported outcomes according to both intention to treat and a per protocol analyses; the intention to treat analysis used the last score carried forward where data was missing. Our reanalysis was restricted to the per protocol groupings since only these individuals had the complete baseline and six-month follow up SDQ scores (n=86) that are required to calculate the Added Value and change scores. As this analysis aimed to evaluate how accurately the SDQ Added Value Score

could predict the effect size obtained by the per protocol analysis in the trial, the attrition biases inherent in per protocol analyses are likely to be irrelevant. For the purposes of this paper, we were interested in whether the SDQ Added Value Score could reflect the effect of treatment as reported, rather than estimating the true effect of the trial intervention adjusting for dropouts.

The intervention in the original trial was highly effective according to the primary outcome measure (ECBI problem scale; effect size = 0.70, 95% confidence interval 0.33 to 1.06), with weaker effects according to the more general SDQ (effect size = 0.37, 0.005 to 0.73 according to SDQ total difficulties score). These effect sizes were calculated from analysis of covariance of the response, taking account of area, treatment and baseline measurement.

Statistical analysis

The analysis was conducted using SPSS for Windows 15.0. The Added Value Scores and change scores were calculated for each child using the equations below.

Raw SDQ Added Value Score (in SDQ points) = 2.3 + 0.8*baseline total difficulties score + 0.2*baseline impact score – 0.3*baseline emotional difficulties subscale score – follow-up total difficulties score.

Raw change score (in SDQ points) = baseline total difficulties score – follow-up total difficulties score

Effect sizes were calculated from the raw scores for the both Added Value and change scores by dividing the raw scores by their respective standard deviations in normative samples (5.8 for the total difficulties score, 5 for the Added Value Score; see www.sdqinfo.com). If the algorithm for the SDQ Added Value Score worked as we expected, the Added Value Score for the control group should be zero (i.e. no change as no intervention), and the Added Value Score for the intervention group should approximate to the effect size reported in the original trial (0.37). We tested whether the

observed mean effect sizes for the SDQ Added Value Score and simple change scores differed significantly from the expected values in the intervention arm (that reported by the trial) and the control arm (no effect as no intervention) using a one-sample t test. The one sample t-test compared the mean of the experimental sample (i.e. the SDQ Added Value Score or the change scores) with a comparison mean set to the expected value for each group (i.e.0.37 for the intervention group and 0 for the control group).

Results

As Table 1 illustrates, the sample from which the SDQ Added Value Score was derived and evaluated resembled the Sure Start sample in gender but differed markedly from it in mean age and more modestly in the mean level of emotional and behavioural difficulties. If the SDQ Added Value Score failed to predict the impact of the intervention as predicted, we would not know if this was because the algorithm did not work, or because the context was so different. However, if the SDQ Added Value Score functioned as expected, these differences would provide evidence for the algorithm's robustness to contextual change, in line with the weak relationship between complexity factors and the Added value Score in the sample from which it was derived. By comparison with the rest of the British Child and Adolescent Mental Health Survey 2004, the SDQ Added Value Score derivation sample was slightly older, more often male, and had a much higher level of emotional and behavioural difficulties; as would be expected for a subsample designed to resemble the sorts of children seen by mental health clinics.

Insert table 1 about here

As shown in Table 2, the effect size based on the Added Value Score of the control group was very close to zero (-0.03, -0.30 to 0.24), which is the *a priori* predicted value for a group who received no treatment. By contrast, the effect size based on simple change scores for the control group was 0.35 (0.12 to 0.59), presumably indicating the failure to account for regression to the mean, attenuation and spontaneous improvement. Likewise, the effect size for the Added Value Score of the intervention group was very close to effect size reported in the original trial (trial 0.37; 0.005 to 0.73; Added Value Score

0.36; 0.12 to 0.96). The effect size for the change score among the intervention group was 0.65 (0.43 to 0.87), representing a considerable overestimate of the impact of the intervention in the trial as assessed by the SDQ total difficulties score.

Insert Table 2 about here

The effect sizes calculated from the Added Value Score were not significantly different to the expected values for either arm of the trial (intervention t=-0.1, p=0.9; control t=-0.2, p=0.8), while the effect sizes derived from the change scores were significantly different to the expected values in both the intervention (t=2.5, p=0.01), and control (t=2.9, p=0.005) groups.

Discussion

Substantive findings

The SDQ Added Value Score behaved as predicted by producing an effect size close to zero for the control group and an effect size for the intervention group that was virtually identical to that calculated using SDQ total difficulties scores in the original trial. By contrast, simple change scores suggested a substantial impact from being on a waiting list in the control group, and also considerably overestimated the effectiveness of the intervention. These findings provide preliminary support for the use of the SDQ Added Value Score to assess the effectiveness of interventions with children who have, or are at high risk of, impairing psychopathology. This is reassuring since a public service agreement based on the SDQ Added Value Score has provisionally been recommended for adoption in England in 2009.[18] Nevertheless, we have only validated the Added Value Score by reanalyzing a single trial, and further replication is a priority.

Only a very small proportion of the variance of the SDQ added value score was explained by the baseline characteristics of the children participating in the British child mental health survey 2004, which is not surprising given that case complexity measures based on factors theoretically important to the outcome of child mental health interventions are not closely related to outcome when studied in routine clinical services.[19] However, concerns

about the difficulty in measuring case complexity and case-mix remain a major impediment to routine outcome monitoring.[20] It is possible that the SDQ added value score might be influenced by characteristics that were not measured in the baseline survey, but those factors commonly thought to contribute to case complexity in child mental health were examined. It may be that case complexity adds to practitioner work load in child mental health services, in terms of the number of professionals involved, the number of appointments offered and the increased liaison required with multiple agencies, but that more complex cases do not inevitably have a worse outcome. This would explain practitioners concerns about case complexity and is an important empirical question for those involved in service development.

Limitations

53% of the children from the Incredible Years trial were three years of age, while the version of the SDQ used is aimed at 4-16 year olds.[12] Younger children are likely to exhibit argumentative or disobedient behaviour rather than more severe difficulties tapped by some questions in the school aged version of the SDQ (e.g lying or stealing.) It may have underestimated behaviour problems and any subsequent change. However, these two versions of the SDQ are identical except for the substitution of two items relating to oppositionality for the conduct disorder questions, and the softening of one item relating to overactivity and inattention in the 3-4 year old version, so that any underestimate in a high risk sample is likely to be small. More importantly in relation to the current study, an underestimate in the level of behaviour problems is immaterial as long as there was a statistically significant difference between the intervention and control arms according to the SDQ that would allow us to test the algorithm. The important issue was whether the Added Value Score could replicate the SDQ effect size estimated by means of a randomized controlled trial (the "gold standard"). That the SDQ added value score produced results so similar to the trial in 3-4 years olds despite being derived on an older population (5-16) provides some evidence that the algorithm can work in populations other than that from which it was derived.

As the Incredible Years randomized controlled trial did not use the follow up version of the SDQ, we were unable to examine how the Added Value Score compared to the responses of parents in the trial sample to the additional questions in the follow up SDQ about whether their child's difficulties had improved or whether the intervention had helped in other ways. We were only able to investigate this source of face validity in the sample from which the algorithm was derived, with obvious limitations. The only difference between the follow up and ordinary versions of the SDQ is the time period that the informant is asked about; one month and six months respectively. The shorter time period at follow up is thought to allow time for the intervention to have an impact and to focus the informant's mind on more recent functioning. The longer time period used in the trial may have diminished the difference between the trial and intervention groups, but as stated above, the key test for the algorithm was whether it could replicate the findings of the trial, rather than precise estimation of the effectiveness of the intervention.

Clinical and policy applications

The original trial reported a larger effect size (0.70, 0.33 to 1.06) according to the Eyberg Child Behaviour Inventory, which is a specific measure of behavioural difficulties that is designed for 2-16 year olds, than with the more broadly focused SDQ (0.37, 0.005 to 0.73). This illustrates a recognized tendency for broad outcome measures to produce smaller effect sizes than specialized measures.[21] Such effects needs to be acknowledged when broad outcome measures are used in routine outcome monitoring so that low effect sizes do not inappropriately discourage practitioners and their commissioners. While the SDQ has the advantage of allowing comparison across children with disparate problems and access to general population norms, clinicians may want to supplement routine monitoring of the outcome of all cases with the SDQ with disorder-specific scales.

The fact that the SDQ is a broad-focus measure is one reason why it is unrealistic to expect CAMHS practitioners in everyday practice to replicate the effect sizes of 0.5 or greater which are often reported in efficacy trials using specialized measures that focus on the problem being treated. In addition, efficacy studies typically involve children without comorbid difficulties, and results for such children do not necessarily translate

easily to children attending mental health services, where comorbidity is the rule.[20,22] Other important caveats for the appropriate use of the Added Value Score are set out in Box 1.

As the formula was derived from children who had psychiatric disorders or whose parents were concerned about their child's mental health, both of which reduced the level of spontaneous improvement over the subsequent six months, the SDQ Added Value Score will *underestimate* the level of spontaneous improvement and thus *overestimate* the impact of any intervention if applied to children with milder problems. It is therefore *inappropriate* to use the current algorithm to assess primary prevention or interventions among children with low levels of initial difficulty. While the confidence intervals around the scores of an individual child are too wide for the SDQ Added Value Score to be a reliable index of that child's progress, our findings suggest that for groups of children treated by a clinician, team or clinic, it can detect significant change. Examination of responses to the SDQ at baseline and follow up may help case formulations or clinical discussions on an individual level.

The concept of clinically significant change, defined as a statistically reliable return to normal functioning, and the related reliable change index have been proposed as tools for evaluating the impact of psychological interventions.[23] However, the cut points denoting clinical significance are inevitably arbitrary, a return to normal function is not expected in many children (autism for instance), and this approach may not be appropriate for patients with comorbid problems (most of those attending child mental health services).[23,24] As the SDQ Added Value Score relies heavily on the impact scores at baseline, it detects therapeutic impact on function as well as symptoms, and is not constrained by comorbidity or where a return to normal function is not feasible. In addition, it uses a quasi-experimental comparison group, rather than essentially arbitrary cut points to assess clinical significance. The mean level of symptoms in a population is related to the prevalence of psychological distress in that population, and the "normal" level of symptoms or impairment among children is not known.

Lambert and colleagues (2007) have used a huge database of responses to one particular questionnaire to provide feedback to therapists about how adult service users are responding to treatment.[25] The questionnaire is completed prior to each session and therapists provided with feedback produce better results among patients who are not responding or deteriorating than therapists who do not receive this advice. They have developed a measure for children and young people, but they have yet to establish its psychometric properties, there is not yet a large database to base practice on, and although promising, this method is dependent on clinically significant change calculations, with all the difficulties discussed above.

A recent review suggests that the publication of outcome data stimulates quality improvement activity; although the papers included were dominated by cardiac surgery and there was inconsistent evidence of improved effectiveness.[26] Australia leads the world in routine outcome monitoring in mental health, including CAMHS, and in adults has been able to demonstrate the effectiveness of mental health services from centrally collated mandatory data (see www.mhnocc.org).[27]

Even if demonstrated to be reliable with repeated testing, the SDQ Added Value Score is just one tool for assessing the quality of services. For the best assessment of service provision and development, service should collect a combination of measures such as clinician and service user rated questionnaires on outcome, satisfaction reported by parents and young people, direct observational measures and process measures. The best assessment of quality will be achieved by triangulating data from different sources and looking for explanations for both good and poor results. As the follow up study used to generate the Added Value Score only collected SDQ's from parents, there are not yet equivalent Added Value Scores measuring the impact of interventions as reported by teachers or young people themselves.

As Lilford states, (2007), the emphasis in outcome monitoring should be on encouraging improvements by all rather than seeking to "name and shame" those who have poor results in some areas: most services will have a spectrum of results.[20] Ranking services or

measuring them against an average measure is certain to undermine morale, because someone has to be the "worst" and by the laws of statistics approximately half will be "below average". Moreover, such an exercise fails to inform us about the absolute quality of the services provided; one service will still be ranked lowest, even if every service exceeded every performance target set.

A recent comparison of hospital episode statistics and the central cardiac audit database suggested that incomplete and/or inaccurate data can lead to highly misleading findings; which if placed in the policy or public domain, can have highly adverse impacts on services.[28] Complete and accurate data is therefore crucial, and most services will need additional resources to develop high quality data management programmes with universal procedures for entry and regular auditing.[28] Only in this way will we be able to draw reliable conclusions about what works for improving child mental health in routine clinical practice.

The SDQ Added Value Score is an outcome-based measure of CAMHS quality. Lilford and colleagues argue that measures of process are preferable to outcome measures, in that process measures are less likely to create perverse incentives and are better correlated with quality.[20] While we strongly agree that it is important to reflect on the process and content of care, we do not believe that all outcome measures should necessarily be excluded from quality evaluations. The SDQ measures the type of difficulties that lead families to seek help and their impact, which are legitimate targets of intervention. The SDQ Added Value Score seems to be relatively robust to the complexity factors which Lilford *et al.* argue will tend to influence many outcome measures. Being completed by parents, the SDQ added value score is less vulnerable than clinician rated measures to gaming to meet management targets, and arguably less likely to create perverse incentives.[20] It is also important to remember that child mental health is one area where we actually have relatively limited data as to which 'processes' *do* improve child mental health when delivered in routine clinical settings. We therefore believe that, if the encouraging findings from this first evaluation can be replicated, then the SDQ Added Value Score may prove an important tool for evaluating CAMHS quality.

Contributors

RG conceived the idea of an Added Value Score, and this was developed by AG. TF conceived the idea of testing it against data from a randomised control trial, conducted the analysis with RG and took the lead in writing the paper. JH and TB designed and led the RCT of parent training and AG suggested testing the Added Value Score against simple change scores. All authors contributed to the writing of the manuscript. TF is the guarantor of for the study.

Declaration of Interests

RG and AG are directors and part owners of Youthinmind, which provides the www.sdqinfo.com website as a public service in order to make the SDQ freely available in many languages for non-profit use and to publish SDQ norms and the Added Value Score formula. All the other authors have no competing interests.

Acknowledgements

RG and TF's membership of the CAMHS Outcome Research Consortium (see www.corc.net.uk; a collaboration of mental health services, academics and policy advisers who are working on an outcome monitoring protocol) stimulated them to design and evaluate the SDQ Added Value Score.

The British Child and Adolescent Mental Health Survey 2004 was funded by the Department of Health; the Health Foundation funded the trial of parent training and TF wrote this paper while supported on an MRC clinician scientist fellowship.; none of these funders had any involvement in the design or analysis of this paper or the construction of the added value score

References

1. Weisz JR, Jensen AL. Child and adolescent psychotherapy in research and practice contexts: review of the evidence and suggestions for improving the field. *Eur Child Adolesc Psychiatry* 2001;**10, Supplement**:12-18.

2. Department of Health. *Health Service Circular 2003/003, Local Authority circular (2003)2. Child and Adolescent Mental Health Service grant guidance.* www.doh.gov.uk/publications/coinh.html.

3. Bridgewater, A., Grayson, A., Brooks, N., Grotte, G., Fabri, B., Au, J., Hooper, T., Jones, M., Keogh, B. Has the publication of cardiac surgery outcome data been associated with changes in practice in Northwest England? An analysis of 25,730 patients undergoing CABG surgery under 30 surgeons over 8 years. *Heart* 2007;**93**:744-748.

4. Department of Health. *Getting the right start: national framework for children. Emerging findings.* London: TSO 2003.

5. Last JM. (Ed). *A dictionary of Epidemiology; third edition.* New York.: Oxford University Press 1995 page 144.

6. Jensen PS, Roper M, Fisher P, Piacentini J, Canino G, Richters J, Rubio-Stipec M, Dulkan MK, Goodman S, Davies M, Rae D, Shaffer D, Bird HR, Lahey BB, Schwab-Stone ME.). Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). *Arch Gen Psychiatry* 1995;**52**:61-71.

7. Ford T, Collishaw S, Meltzer H, Goodman R. A prospective study of childhood psychopathology; predictors of change over three year. *Soc Psychiatry Psychiatric Epidem* 2007;**42**:953-961.

8. Cole T, Flegal KM, Nicholls D, Jackson AA.. Body Mass Index cut offs to define thinness in children and adolescents. *BMJ* 2007;**335**:194-197.

9. Cotterill AM, Majrowski WH, Hearn S, Preece MA, Savage MA. The potential effect of the UK 1990 height centile charts on community growth surveillance. *Arch Dis Child* 1996;**74**:452-4.

10. Green, H., McGinnity, A., Meltzer, H., Ford, T, Goodman, R. *Mental health of children and young people in Great Britain, 2004*. London: TSO 2005.:

11. Hutchings J, Bywater T, Daley D, Gardner F, Whitaker C, Jones K, Eames C, Edward, R. Parenting interventions in Sure Start for children at risk of developing conduct disorder; pragmatic randomised controlled trial. *BMJ* 2007;**334**:678-82.

12. Goodman, R. Psychometric properties of the Strengths and Difficulties Questionnaire (SDQ). *J Am Acad Child Adolesc Psychiatry* 2001;**40**:1337-1345.

13. Goodman R, Ford T, Richards H, Meltzer, H, Gatward, R. The Development and Well-being Assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *J Child Psychol Psychiatry* 2000;**41**:645-657.

14. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders; Diagnostic Criteria for Research.* Geneva: World Health Organization 1993.

15. Webster-Stratton C. Preventing conduct problems in Head Start children: strengthening parenting competencies. *J Consult Clin Psychol* 1998;**66**:715-730.

16. NICE. *Parent training / Education programmes in the management of children with conduct disorders. NICE technology appraisal guidance 102*. London: SCIE, NHS. 2006. Available from www.nice.org.uk/TA102.

17. Eyberg S, Ross AW. Assessment of child behaviour problems; the validation of a new inventory. *Journal of Clinical Child Psychology* 1978;**7**:113-116.

18. HM Treasury. *PSA Delivery Agreement 12: improve the health and well-being of children and adolescents*. London:.TSO 2007.

19. Garralda ME, Yates P & Higginson I. (2000). Child and adolescent mental health service use: HONOSCA as an outcome measure. *British Journal of Psychiatry 177,* 52-58.

20. Lilford RJ, Brown CA, Nicholl J. Use of process measures to monitor the quality of care. *BMJ* 2007;**335**:648-650.

21. Lee W, Jones L, Goodman R, Heyman I. Broad outcome measures may underestimate effectiveness; an instrument comparison survey. *Child Adolesc Mental Health* 2005;**10**:143-144.

22. Ford T, Hamilton H, Meltzer H, Goodman R. Child mental health is everybody's business; the prevalence of contacts with public sectors services by the types of disorder among British school children in a three-year period. *Child Adolesc Mental Health* 2007;**12**:13-20.

23. Jacobson NS, Roberts, LJ, Berns SB & McGlinchey JB. (1999). Methods for defining and determiniming the clinical significance of treatment effects: description, application and alternatives. *Journal of Consulting and Clinical Psychology* **67** 300-307.

24. Wise, EA. Methods for analyzing psychotherapy outcomes:a review of clinical significance, reliable change and recommendations for future directions. *Journal of Personality Assessment* **82:** 50-59.

25. Lambert M. (2007). Presidential address: what we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research* **17:** 1-14.

26. Fung CH, Lim YW, Mattke S, Damberg C& Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Annals Intern Med* 2008; **148**:111-123.

27. Burgess P, Pirkis J & Coombs T. (2006). Do adults in contact with Australia's public sector mental health services get better? *Australia and New Zealand Health Policy* **3:** 9-16.

28. Westaby S, Archer N, Manning N, Adwani S, Grebnik C, Ormerod O, Pillai R, Wilson N. Comparison of hospital episode statistics and central cardiac audit database in public reporting of congenital heart surgery and mortality. *BMJ* 2007;**335**:759-762.

Table 1 Comparison of the samples from which the SDQ Added Value Score was derived and evaluated (Welsh Sure Start Trial).

		British Child and Adolescent Mental Health Survey 2004 (n=7977)[1]	SDQ Added value score derivation sample (n=609) [1]	Welsh Sure Start Trial (n=133)
Age in years	Range	5-16	5-16	3-4
	Mean (standard deviation)	10.5*** (3.4)	11.0 (3.3)	3.9*** (0.5)
Male gender (%)		51.5***	61.1	60.2
Mean SDQ parental total difficulties score at baseline (standard deviation)		7.9*** (5.9)	15.5 (7.2)	17.7*** (5.8)

1. SDQ added value score derivation sample is a sub-sample of the British Child and Adolescent Mental Health Survey sample 2. Chi-Square and t-tests use the SDQ derivation sample as the reference group for comparison with the remainder of the British Child and Adolescent Survey and with the Welsh Sure Start Trial: ***=p<0.001

Table 2 Comparison of the Added Value SDQ scores and change scores with the expected effect sizes for control and intervention groups separately

SDQ Added Value Score = expected SDQ total difficulties score at follow up- observed SDQ total difficulties at follow up

Change scores = baseline total difficulties score- follow-up total difficulties score.

	Effect size in standard deviation units (95% confidence interval)		
	Expected value[1]	Added Value Score	Change score
Control group (n=47)	0	-0.03 (-0.30-0.24)	0.35 (0.12 - 0.59)*
Intervention group (n=86)	0.37 (0.005-0.73)	0.36 (012-0.60)	0.65 (0.43 - 0.87)**

1 The expected value for the control group was predicted as 0 a priori, because they received no treatment, while the expected value for the intervention was the effect size reported from the original trial according to the SDQ. * p<0.05, **p<0.01 value significantly different to that expected.

Box 1 Caveats for clinical practice

- The added value score is only calibrated for use with therapeutic or targeted interventions and will overestimate change in groups with low levels of psychopathology. It should *not* be applied to universal interventions.

- The added value score is a tool for evaluating the impact of interventions on *groups* of children, and the confidence intervals around the scores of *individual* children will be too wide to interpret in most instances.

- The added value score requires follow up to occur between 4 and 8 months after the initial measure. Follow up after a fixed interval is preferable to administration at discharge because of the risk that discharge may follow soon after a spontaneous improvement, and thereby capitalize on chance remission.

- The added value score is based on the SDQ, which is a "wide angle" measure. Clinicians may want to supplement the SDQ with more specific outcome measures relating to each child's individual problems.

- The use of multiple measures (clinician, parent, child, process, satisfaction, direct observation) will provide commissioners, practitioners and policy makers with richer data for improving services

- Services need to aim for high response rates from parents in order to obtain representative data. This requires resources.

Figure 1 Mean SDQ Added Value Score and 95% confidence intervals in relation to parent's opinion about their child's difficulties at follow up in the sample from which the algorithm was derived