Research Section

Goal Attainment Scaling: A Technique for Evaluating Conductive Education

Gilbert MacKay, Susan McCool, Sally Cheseldine, Elspeth McCartney

Conductive education – how to evaluate it?

Is it really impossible for scientists to design the definitive evaluation study? This is a question asked many times by both parents and professionals involved with conductive education. Understandably frustrated by years of controversy over an approach which was first introduced into this country more than 30 years ago, many looked to the Birmingham project for the answer to the question: is conductive education a good way of educating children with cerebral palsy? Here was a project which had been planned by experts in the field, was funded by the government and was centred on an attempt to *transplant* conductive education as a complete system into the UK. However, as Bairstow and Cochrane pointed out recently in the *BJSE*, the teething problems associated with getting the Birmingham Institute running had a knock-on effect on the evaluation itself and the results raised many more questions than they answered. Ideally, of course, the first Birmingham evaluation should have been viewed as a preparation for a second study which would begin after the first set of conductors had been fully trained and new children recruited.

Since it is unlikely that a repeat of the Birmingham project will take place, what are the alternatives? In the papers that follow, two quite different approaches are represented. In the first MacKay and colleagues, from the Faculty of Education Strathclyde University, describe the beginnings of another large project, based at the Scottish Centre for Children with Motor Impairments. In this project, organ transplant has not been considered. Instead, an attempt is being made to produce a Scottish version of conductive education by the process of grafting. As part of the evaluation project the group is experimenting with a measurement technique, Goal Attainment Scaling, which they describe in detail.

In the second paper, Sigafoos *et al.*, from the Fred and Eleanor Schonell Special Education Research Centre, University of Queensland, Australia, take another approach. Rejecting the notion that 'the whole is more than the sum of the parts' they assume the examination of component elements of conductive education is worthwhile and have done a small scale study of a short term intervention programme. Although purists might argue that neither of these approaches will answer the question 'Does conductive education work?' surely the realists among us will concede that pursuit of the Holy Grail must sometimes give way to more practical projects.

Context

The goal of the Craighalbert evaluation project is to consider the place of the Scottish Centre for Children with Motor Impairments within the context of Scottish educational provision as a whole. As such the project is not concerned with the evaluation of conductive education *per se* but with a system which has strong roots in Scottish under-five education as well as being influenced by the Hungarian model. Although analysis of the whole system extends beyond evaluating the effectiveness of the service for the children who use it, such assessment is a critical aspect and will inevitably inform the development of future policy. As one component of this assessment process, goal attainment scaling was chosen as a technique as it is specifically designed to accommodate the heterogeneity that exists among children with movement difficulties and the variation in rates of progress that they exhibit. The paper provides an introduction to the technique and some of our initial reactions to the experience of using it.

Background to the problem

The evaluation of services designed for small groups of atypical individuals presents many challenges to researchers. Whether they be Olympic athletes or people with severe intellectual difficulties, one of the most difficult problems by far is the selection of techniques and instruments which can be used to determine whether a particular kind of intervention is having its intended effect. By definition, such groups are not representative of the general population and therefore comparing their characteristics with those of the general population can cause problems. In addition, the variability that exists within such groups, irrespective of any single label under which they are classified, makes it difficult to apply the same measures to all members at any one point in time. Because of this variability, too, it is usually impossible to create subgroups which could be compared with any confidence.

This paper has been written to make a case for the use of goal attainment scaling, a technique which has been used in service evaluation in the United States since the late 1960s but which has been used only rarely in the UK (although see Stanley, 1984, and Imich and Roberts, 1990, for exceptions). As an alternative or adjunct to the more traditional methodologies used goal attainment scaling seems to have much to offer, precisely because it was designed to deal with the sorts of difficulties alluded to above. To put the method in context, it may be helpful to begin by referring very briefly to two of the more traditional approaches used in evaluation studies – the use of norm referenced tests and multiple base-lining.

Normative tests

Since the turn of the century, the assessment of children by reference to the norms of large, representative populations has been common practice and certainly has its place in the evaluation of services. For example, although the data may not be so useful to the practitioners providing the service, policy makers may find it helpful to establish how a particular group compares to the general population. However, for those concerned with documenting the progress of individuals within a programme of intervention or treatment, such measures can be problematic. For example, some norm referenced tests are difficult to use with individuals at the extreme ends of the normal distribution – the tasks within them may be either too difficult or too easy. Other tests can be used but do not offer scoring systems which are sensitive to small differences between individuals or to small changes over time. Yet others do not seem to focus on the correct aspects of the behaviours being investigated. For example, there are tests of communication which do not come close to confronting the spontaneity and diversity that characterise communication in everyday life.

The difficulties associated with using norm referenced tests increase even more when services exist to help clients develop characteristics which are complex conceptually, such as 'effective communication', 'functional movement' or 'independence'. Since few norm referenced tests have been designed to measure attributes at this level, the measures they yield are often incompatible with the more global measure required. For example, whereas a measure of 'functional movement' might give a physically impaired child a high rating if he can walk a set distance – irrespective of how he does it – reference to the 'norm' might result in the same child being given a low rating. Even when instruments exist which attempt to measure these global characteristics, however, they often share the problems of the norm referenced test in that the scales

Table 1 Sample goals scaled for an individual child

	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	
+2	Cross mid-line	Spontaneously grasps and releases object	Begins to initiate dressing activity, such as pulling up pants	Initiates helping with putting on shirt	Puts coloured blocks in container by colour on command	Best likely outcome
+1	Uses hands at mid-line for 1 minute	Grasps and releases object on command 4 times during test session	Removes shoes and socks	Holds head steady for putting on shirt (no reward)	Puts blocks in container, occasionally matching colour to adult's command	Better than expected
0	Holds 6" ball with both hands for 10 seconds	Initiates release of object	Removes both socks	Holds up head for putting on shirt voluntarily, for reward	Places blocks in container purposefully using colour to organise them	Goal
-1	Hands to mid-line, but unable to do functional tasks	Cannot release object (X)	Needs assistance to start removing socks	Occasionally will hold up head (for reward)	Random placement of blocks in container (X)	Disappointing
-2	Cannot get hands to midline (X)	Ignores objects and does not try to grasp and release	Does not start to try removing socks on request (X)	Does not hold head up (X)	Ignores container	Worst likely outcome
(Weighting)	2	1	1	10	1	10

within them are crude, have too few steps and are difficult to score outside a limited range.

Finally, the use of norm referenced tests is problematic if one wishes to characterise changes in a group of children in relation to the specific objectives of the programme. Many services exist which are designed for groups that are superficially homogeneous, such as pre-school children with cerebral palsy. As most practitioners will know the variation that will exist within any such group is likely to be so great that the use of any kind of norm referenced test to characterise group change would be likely to be uninformative. In short, although normative measures have a role to play in evaluation, it is generally accepted that they need to be supplemented with measures that are designed specifically to measure change as it is defined within the particular intervention being investigated.

Multiple base-lining

Multiple base-lining became a popular technique in the 1970s for evaluating the progress of children with developmental difficulties and adults with intellectual difficulties. It involves, first, establishing a client's base-line scores in a collection of areas, for example, communication, motor performance, self help skills, social skills and cognition. A period of intervention follows in just one of these areas and then the original measures are repeated. The simplest hypothesis checked by this procedure is that the client will improve in the specific area of intervention only. Of course, more complex issues may emerge when results appear. The technique is also appropriate for evaluating progress in areas for which there are no ready-made scales; investigators can tailor their own, to meet special circumstances and criteria.

One advantage of multiple base-lining over traditional methodologies is that the clients of services act as their own controls. This overcomes problems which arise when attempting to compare groups or determine the significance of improvements along developmental lines. The strength of multiple base-lining lies in investigating single characteristics which a programme of intervention is intended to develop. The development of intentional behaviour through conductive education is one possible example which the authors may investigate in their current project. However, the technique does not cope so well when a service is designed to achieve a diverse range of outcomes by a united approach. That is where goal attainment scaling seems to have something to offer.

Goal Attainment Scaling

As noted above, goal attainment scaling was first developed by American health agencies more than 20 years ago (Kiresuk and Sherman, 1968) and has been used frequently since (*eg* Maloney, Mirret, Brookes, and Johannes, 1978; Carr, 1979; Mayer, 1983; Bailey and Simeonsson, 1988), though its use in Britian is rare. Put at its simplest, the process of goal attainment scaling is based on the setting of a number of goals for each individual within the service being provided and the measurement of progress in relation to these goals. More specifically, the process begins with the service providers setting around five goals, each of which becomes the mid-point of a five-point scale (see below for an example). Progress from base line points is usually measured at three-monthly intervals. Since the setting of these goals is done entirely on an individual basis, any problems associated with the heterogeneity of the individuals within the group is avoided. Each child or client has his or her own goals, own base-points and progress markers and is treated as an individual. However, since the method *also* includes statistical procedures for standardising the individual's scores, it is possible to compare individuals within the group too.

The need for training in goal setting is an obvious factor in the success of any such project. Many professionals who have been conversant with similar procedures, such as those basic to the 'objectives' movement, will find the technique straightforward. Others might need considerable help (Choate, Smith, Cardillo, and Thompson, 1981). In addition to the professional background and previous experience of the staff involved, however, knowledge of the way competence develops in the required domains might also be helpful. At present, the selection of goals and steps within goals sometimes seems to take place in a theoretical vacuum. This can result in minor problems such as steps being inappropriately spaced or even major problems relating to the internal coherence of the scales that emanate from the central goal. For instance, in the example of 'removing socks' given in Table 1 it is difficult to see how the item described as the 'best likely outcome' relates to the scale it is in.

An example of the technique in practice

In the Craighalbert study, members of the research team have worked with staff to address questions about the compiling of scales and to ensure that the intended goals of intervention can be observed as objectively as possible. An example of the outcome of one of these group meetings for a given child is given in Table 1.

As can be seen from the table the process begins with the specification of five goals which the staff who work with the child think are important to attain within, say, a school term. Each intended goal is placed in the centre of a five point scale and takes the value of 0. In the light of their knowledge of the child, the staff then specify attainments which would represent the worst possible outcome and assign this the value of −2, the best possible outcome, assigned the value +2, disappointing −1; and better than expected +1. One of these values is then recorded as the child's current level of performance and changes in relation to that are noted after the requisite three-month period has elapsed.

In addition to considering each of the five goals for a child as individual objectives, it is common practice to weight goals to ensure that the eventual standard score is biased towards those that the staff consider to matter most. This is done by assigning weights that are agreed in discussion among the staff team. For example, as Table 1 shows, a team has decided that goal 1 is twice as important as goals 2, 3 and 5 and that goal 4 is 10 times as important as goal 1. Of course, goals do not have to be weighted and, in that event, are given a weighting of 1 for the purposes of calculating standard scores.

The five final scores for each item, together with any weightings assigned to them, are translated into a single standard score using the formula described by Kiresuk and Sherman (1968, p.355-356). Stanley (1984) provides a helpful worked example of the calculation.

In January 1993, we gathered data from our first attempts to use goal attainment scaling for a group of 14 children attending the Scottish Centre for Children with Motor Impairments and on the whole our reaction is positive. Although it is not yet possible to report on the results of this investigation out of respect for the conditions that apply to the research grant, using the standard scores for these 14 children we are in the process of seeking empirical answers to a number of simple questions. For instance, are there significant differences between the overall outcome scores for subgroups, such as boys and girls? Is there any pattern in the distribution of attainment scores – are they evenly distributed around the goal or are they skewed in such a way that over or under-estimations by the professionals might be detected? At present answers to such questions are not of great importance as this is the first real-life exposure to goal attainment scaling that both the staff of the centre and the evaluation team have had. What is important, however, is that a set of procedures for setting goals is now in place, outcome measures on a three-monthly basis can be obtained and standard scores for these outcomes can be calculated. However, ease of obtaining results is no guarantee that they will be reliable or valid and thus it is important to heed the concerns expressed by Cytrynbaum, Ginath, Birdwell, and Brandt (1979) about cavalier use of the technique without continual discussion and refinement.

Since we are keen to collaborate with other colleagues in any discipline where goal attainment scaling is being used, we felt it might be useful to complete this paper by listing what we think are the advantages of the technique and conclude by pointing to a number of problems we have already considered.

Advantages of the technique

Goal attainment scaling (GAS) is a particularly appropriate evaluation tool for the following reasons.

1 The goals for the individual are tailored to that individual's personal needs as they are catered for within the service/programme. Thus, there is no need to refer to conventional, published charts and scales of development which may have little immediate relevance. This is particularly appropriate in the context of conductive education, where the goals set are phrased in terms of the global concept of 'orthofunction', a 'capacity . . . enabling an individual to satisfy . . . biological and social demands' (Hári and Akos, 1988. p.140).

2 Statistically derived standard scores can also be assigned to the performance of individual subjects, making it possible to compare the scores of individuals within a group. Also, data expressed in this way may then be analysed in relation to any other relevant factor such as the clinical diagnosis of clients, the distance they travel to the service, characteristics of the staff and families, and so on. When dealing with small heterogeneous groups, this is very difficult using more traditional assessment procedures.

3 It is possible to make a qualitative analysis of goals and progress towards them to discover the contexts to which they relate. For example, one might ask how the goals set for five-year-olds relate to national expectations as they are expressed in curriculum content. Similarly, in our own project we might inquire how the goals set for individual two- and three-year-olds are reiated to principles of programme development in pre-five education. On a broader scale one might also inquire how goals relate to the idea of valued life in the community.

4 GAS is able to cope with the unique methodologies and claims of approaches that may otherwise seem to defy evaluation by more traditional methods. In practical terms, what this means is that practitioners can set their goals in any terms they wish provided it satisfies their philosophy and goals are specified in a way that makes reliable observation possible.

5 GAS can also be used to chart the development of a service historically, in that it allows analysis of the way goal setting changes and of how well goals are attained over time.

The results of GAS complement data from conventional, normative measures in that they help to create a clearer context for understanding the results of such measures especially when they are applied to 'extreme' populations. Also, such data can be used to reveal change when normative measures seem insufficiently sensitive to small changes in a child's behaviour in relation to particular goals.

Some problems with goal attainment scaling

So far the problems we have encountered with goal attainment scaling relate mainly to the statistical procedures that are recommended to deal with the data obtained in practice.

To begin with, we have some difficulty with the calculation of standard scores and the way the scales are treated. The goal attainment scale for an individual subject is a standard score obtained by calculating standard deviations from the 'expected' target score of 0. We have followed the customary GAS procedure of translating children's goal scores into standard scores with a mean of 50 and a standard deviation of 10. The translation formula chosen was:

$$SS = 50 + \frac{10\left[\sum w_i x_i - E\left(\sum w_i x_i\right)\right]}{\sqrt{VAR\left(\sum w_i x_i\right)}},$$

where x refers to the weightings staff attach to goals, R refers to the actual goals achieved, and E refers to the expected score of 0 (Kiresuk and Sherman, 1968, p.448).

The choice of an expected score of 0 reflects the philosophy of GAS in that all subjects are expected to reach a target score of 0 within a predicted period of time. Therefore, their progress towards goals and the ability of intervention teams to predict and set goals may be judged against this expected outcome rather than against some other externally determined and less appropriate standard.

Once one begins to work with these scores and scales, however, problems arise. For instance, each subject's standard score is based on a standard deviation drawn from a set of no more than five raw scores. In addition, these raw scores are treated as *interval* data when it would be equally acceptable, if not more so, to treat them as *ordinal* points on a five-point scale. The conventions of statistics require the exercise of considerable caution when using parametric techniques, on data of this sort.

A further complication exists. The formula above, used currently in the Craighalbert evaluation for calculating standard scores, is a simple variant of a more familiar procedure. However, Kiresuk and Sherman (1968, p.449) recommend the use of an extension of this formula,

$$SS = 50 + \frac{10\sum w_i x_i}{\sqrt{(1-\rho)\sum w_i^2 + \rho\left(\sum w_i\right)^2}}$$

which incorporates a constant, p, that 'bears an *intuitive* (*our emphasis*) meaning of a kind of weighted average correlation' among the goal scores. Later, on p.449, they say that a 'value for p . . . must be assumed. In most cases it will be sufficient to assume a value of, say, p = .3 without formal justification.' These statements seem to beg so many questions that it is difficult to use the extended formula with confidence, despite its long tradition of use in North America.

Table 2 Hypothetical data from goal scales, in parametric and non-parametric form

	Raw scores for each goal					Frequency of attaining score levels A-E				
Child	g^1	g^2	g^3	g^4	g^5	A	B	C	D	E
Anne	0	−2	−1	−2	−1	0	1	0	2	2
Brenda	2	−1	0	0	−1	1	2	0	2	0
Carol	1	1	2	1	0	4	1	0	0	0
Donna	−1	2	1	−2	2	3	0	1	0	1
Emily	−2	2	−1	0	0	1	2	1	1	0

Key

g^1-g^5: goals 1 to 5 of the group set for each child

A: targets exceeded; B: targets met; C: targets not met, but beyond baseline; D: scores remaining at baseline; E: scores regressed from baseline.

Thus, there seem to be several problematic aspects about the statistics of this potentially valuable means of setting targets for intervention and for monitoring progress towards them. To clarify these issues, we have applied the formula to simulated data, and MacKay, Somerville, McCall and Sharp intend to report on this. However, some of the preliminary findings from the simulations are worth noting here briefly.

There are doubts that 0, the goal, is a credible expected score. The assumed variance, 1, is also hard to justify. The value chosen for p does not affect the 'normality' of the distribution of scores from simulated samples though it does affect the range of values which emerges. This could be a useful finding as it relates to establishing the limits of scaled scores for a client group. Yet it would still seem wise to query making a formula more complex by incorporating an 'intuitive' constant, especially when it will be applied to numerically-coded qualitative data that relate to specific individuals and to personalised goals.

Alternative types of analysis

Some of the statistical problems just mentioned may be avoided by treating the goal scores as ordinal data. For example, it is possible to count how often children in a sample have scores at each of the five points in the range from −2 to +2. The data may even be coded in rather different ways, for instance, in terms of how often goals are **A** met and exceeded, **B** met, **C** not met, but are beyond baseline, **D** static at baseline and **E** have regressed from baseline. Table 2 illustrates this type of coding with hypothetical data and with baselines of −1 or −2.

The data to the left of the double bar are the children's sets of five raw scores (between −2 and +2) that are treated as interval (parametric) data for conversion into standard scores. The data to the right of the double bar are based on the same information from the children's records. However, here they appear as they might be recorded in terms of the ordinal (non-parametric) categories, A-E, above. If there is a disadvantage in using the right-hand, non-parametric system, it is that it does not generate standard scores for individuals. However, this may not really be a problem if reservations about the validity of the data and computation of standard scores by the conventional GAS procedure are upheld. The clear advantage of the non-parametric approach is that it allows valid investigations of, for example, sub-group differences without making questionable assumptions about the numerical values used in calculations.

It is worth pointing out that the existence of a set of written goals from a sample of subjects is also potentially useful as a source of qualitative data for empirical analysis. For instance, the collection of goals set for a class of pupils could be coded in terms of curricular areas such as communication, cognition, movement and so on, with the reliability of the coding being assessed by an inter-rater check. Feedback on a content analysis of this sort could become a useful focus for discussion between practitioners and evaluators in action research and in other participant-observer approaches.

Finally, it has to be recognised that asking the staff of a school, centre or other provision to provide information on their goals is itself an act of intervention by researchers (Cytrynbaum *et al.*, 1979, pp.17-18). The setting of termly targets occurs frequently, though not universally, in educational and other provision. Indeed, self-monitoring is likely to become even more firmly established in the culture of schools and other public services as a result of current developments in the area of national standards, demonstrable competence, quality indicators and so on. We acknowledge the cooperation of the Craighalbert staff in agreeing to adopt our system of identifying and scrutinising the goals they set for children. We shall be interested to discover the extent to which it influences the processes of the centre, as well as its value in the assessment of outcomes.

References

Bailey, D.B. and Simeonsson, R.J. (1988) Investigation of use of goal attainment scaling to evaluate individual progress of clients with severe and profound mental retardation. **Mental Retardation, 26,** 5, 289-295.

Cytrynbaum, S., Ginath, Y., Birdwell, J. and Brandt, L. (1979) Goal attainment scaling: a critical review. **Evaluation Quarterly, 3,** 1, 5-40.

Carr, R.A. (1979) Goal attainment scaling as a useful tool for evaluating progress in special education. **Exceptional Children, 46,** 1, 88-95.

Choate, R., Smith, A., Cardillo, J.E. and Thompson, L. (1981) Training in the use of goal attainment scaling. **Community Mental Health Journal, 17,** 2, 171-181.

Hári, M. and Ákos, K. (1988) **Conductive Education** (translation of 1971 text). London: Routledge.

Imich, A. and Roberts, A. (1990) Promoting positive behaviour: an evaluation of a behaviour support project. **Educational Psychology in Practice, 5,** 4, 201-209.

Kiresuk, T.J. and Sherman, R.E. (1968) Goal attainment scaling: a general method for evaluating comprehensive community health programs. **Community Mental Health Journal, 4,** 6, 443-453.

Maloney, F.P., Mirret, P., Brookes, C. and Johannes, K. (1978) Use of goal attainment scaling in the treatment and ongoing evaluation of neurologically handicapped children. **American Journal of Occupational Therapy, 32,** 8, 505-510.

Mayer, C.A. (1983) Goal attainment scaling: a method for evaluating special education services. **Exceptional Children, 49,** 6, 529-536.

Stanley, B. (1984) Evaluation of treatment goals. **Journal of Advanced Nursing, 9,** 4, 351-356.

Notes for Contributors to the Research Section

Three copies of manuscripts should be submitted to Dr Sheila Henderson and Ms Jill Porter, Department of Educational Psychology and Special Educational Needs, Institute of Education, London University, 25 Woburn Square, London WC1H 0AA. The authors should in addition retain a copy for themselves (including tables etc) as the editors do not accept responsibility for loss or damage. Rejected manuscripts will not be returned unless specifically requested. Manuscripts must *not* be submitted simultaneously to another journal.

In preparing manuscripts for submission, authors should use the following guidelines:

1 Ideally, submitted manuscripts should be between 4,000 and 8,000 words including tables and references. Longer articles will be considered only under exceptional circumstances. All articles must be preceded by a brief factual summary of the work, no longer than 200 words.

2 Manuscripts should be typed on one side of the paper only, using double spacing and wide margins. Pages should be numbered consecutively and all tables and figures should be typed on separate sheets and placed at the back of the manuscript (after the reference list). The location of tables and figures in the text should be indicated clearly, *eg* by the instruction 'Insert fig 1 about here'. Any abbreviations should be clearly explained in the text.

3 Authors are urged to pay particular attention to the presentation of references and adhere to the guidelines in the Publication Manual of the American Psychological Association (3rd edition, 1983). Copies are available in most university or college libraries. Anyone who does not have access to this publication may contact the editors for further information.

4 Manuscripts will be sent out blind to two referees so it is important to ensure that authors' names can be removed from the paper. Authors are therefore requested to submit a topsheet which includes the title of the paper, authors' names and main appointments, and the address for correspondence only. The title of the paper should also be included as a heading on the first page of the manuscript.

5 All papers should be written in a concise and easily readable style and demonstrate the writers' familiarity with other publications in the field. The article should contain a clear rationale and explanation for the study, including the design and analysis of the results. Discussion of the relevance for educational provision/practice should be an integral part of the study. The aim of the Research Section is to extend the knowledge of all concerned with the education of children with special needs and very high standards will be applied to research submissions.

Research Section in September issue

It is regretted that the symbols were omitted from the key to three figures on p.111 of the contribution by M.L. Au and P.D. Pumfrey, entitled 'Parents' and Teachers' Expectatiions of Children's Attainments: Match or Mismatch?' The missing symbols were a diamond (representing parents) and a square (representing teachers).