J Rehabil Med 2011; 43: 46–49

ORIGINAL REPORT

AGREEMENT BETWEEN TWO DIFFERENT SCORING PROCEDURES FOR GOAL ATTAINMENT SCALING IS LOW

Thamar J.H. Bovend'Eerdt, PhD[1,2,3], Helen Dawes, PhD[2], Hooshang Izadi, PhD[3] and Derick T. Wade, MD[4]

From the [1]Department of Movement Science, Maastricht University, Maastricht, The Netherlands, [2]School of Life Sciences, [3]School of Technology, Oxford Brookes University and [4]Oxford Centre for Enablement, Oxford, UK

Objective: **To investigate the agreement between a patient's therapist and an independent assessor in scoring goal attainment by a patient.**

Methods: **Data were obtained on hospital patients with neurological disorders participating in a randomized trial. The patients' therapists set 2–4 goals using a goal attainment scaling method. Six weeks later attainment was scored by: (*i*) the treating therapists; and (*ii*) an independent assessor unfamiliar with the patient, using a semi-structured interview method with direct assessment as appropriate.**

Results: **A total of 112 goals in 29 neurological patients were used. The intraclass correlation coefficient ($ICC_{(A,k)}$ = 0.478) and limits of agreement (–1.52 ± 24.54) showed poor agreement between the two scoring procedures. There was no systematic bias.**

Conclusion: **The agreement between the patients' therapists scoring the goals and the independent assessor was low, signifying a large difference between the two scoring procedures. Efforts should be made to improve the reproducibility of goal attainment scaling before it is to be used as an outcome measure in blinded randomized controlled trials.**

Key words: goal attainment scaling; rehabilitation; reproducibility of results.

J Rehabil Med 2011; 43: 46–49

Correspondence address: Thamar Bovend'Eerdt, Department of Human Movement Science, Universiteitssingel 50, NL-6200 MD Maastricht, The Netherlands. E-mail: thamar.bovend-eerdt@maastrichtuniversity.nl

Submitted November 17, 2009; accepted August 30, 2010

INTRODUCTION

Goal attainment scaling is increasingly used in multi-disciplinary rehabilitation (1–3), including in people with neurological conditions (4–10). It is a structured method for evaluating the achievement of goals, first introduced in the 1960s by Kiresuk & Sherman (11) within a mental health service. Goal attainment scaling individualizes the outcome measured for each patient, in contrast to conventional measures that comprise of a standard set of items rated in a standard way. It also allows a standardized score to be calculated (11). The validity and inter-rater reliability of goal attainment scaling in clinical populations has been reported as good (10, 12–17).

Goal attainment scaling is an attractive outcome measure for exploring the effectiveness of interventions in randomized controlled trials (RCTs) because it should be sensitive to change and appropriate for evaluating complex interventions (18). However, in randomized studies an independent assessor who is masked for the subject's allocation should measure outcome, to ensure unbiased measurements. If the assessor is to stay masked and remain independent, the assessor will necessarily be unfamiliar with the subject and be unable to draw on information from treating staff.

In order to explore the utility of goal attainment scaling in RCTs this study investigates the reliability and agreement (19) of goal attainment scoring by the patient's therapist and by an independent masked assessor in the context of a randomized study.

METHODS

This analysis is based on data collected in a study investigating the effectiveness of a 6-week programme of motor imagery in neurological rehabilitation (5) (approved by the Oxfordshire Ethics Committee (07/H0605/84)).

Goal attainment scaling method

This study used a standardized method for writing objective goals (20) derived from earlier descriptions (11, 21). Therapists were taught this method of setting goals in a 1 h workshop. The method starts by listing the patient's wishes, expectations and patient's situation. Then the therapists set valued and achievable goals using the following 4 steps:

- specify the target activity;
- specify support needed;
- quantify performance, and;
- specify the time period to achieve the desired state (in this study it was always 6 weeks).

Combining information from these 4 parts results in an objective goal. Each goal was then weighted both for importance and difficulty, which were ranked on a 3-point scale, ranging from 1 (a little importance/difficult) to 3 (very important/difficult) (20).

Once the goal was set in terms of the performance level expected at a specified time (i.e. the "0" scoring level), 4 more performance levels were specified at the specified time. In this study the current level was always set at level –1 as recommended by some authors (21). Defining the other levels (–2, +1, +2) was easily done by varying one or more of the components discussed above (i.e. support, quantification).

Each patient had two therapists (a physiotherapist and an occupational therapist) and at baseline the patient's therapists created up to 4 individualized goals in conjunction with the patient, and these goals were scaled by the therapist. In this study the time specified for goal

measurement was always 6 weeks, at which time both the treating therapists and the masked assessor scored the goal attainment (within 24 h of each other) without knowledge of the other's rating.

The therapists usually treated the patients regularly (several times each week) and could therefore score the goal achievement of their own goals easily. The therapists were simply asked to complete the rating of the patient at 6 weeks.

The independent assessors were also trained in the goal attainment scaling process, mainly in the process for scoring the outcome. They were asked to score the outcome using a mixture of assessing the activity directly and interviewing the patient, which also depended on the cognitive and communicative abilities of the patients. Patients were included in the study if they were able to understand, remember and execute simple commands (operationally defined as the ability to score positive on the first 3 items of the Sheffield screening test for acquired language disorders (22)). Direct assessment simply involved asking the patient to perform the activity. Interviewing required the independent assessor to establish the patient's actual level of attainment as accurately as possible from information provided by the patient. The assessor was not allowed to consult the patient's therapist or any other clinical staff. The assessor had met the patient only once before at the baseline assessment.

The interview was performed using a semi-structured interview with the patient involving the following 3 steps:

- Ask an open question to let the patient describe how he/she executes the task (e.g. *Can you describe to me how you transfer from your wheelchair to the toilet?*).
- Ask open questions during the patient's explanation to get the patient to elaborate on certain points (e.g. *How and where do you park your wheelchair?*).
- Ask more specific questions to get detailed information on the domains. Special effort was put into trying to "measure" ambiguous terms (e.g. walk outdoors safely). This was done by asking specific task-related questions (e.g. *Do you cross the street on your own?* or *Do you need help stepping down or up kerbs?*).

Within the study data were also collected on diagnosis, time since onset, cognitive function using the Short Orientation Memory and Concentration test (23), general motor function using the Motricity Index (24), mobility using the Rivermead Mobility Index (25), personal activities of daily living using the Barthel Activities of Daily Living (ADL) index (26) (score range 0–20), ability to perform ADL activities using the Nottingham Extended ADL scale (27) and arm motor function using the Action Research Arm Test (28).

Analysis

One summary goal attainment scaling (GAS) score was calculated for each patient, contributed to by up to 4 goals. For each patient a total score was calculated by applying the usual formula (11, 21):

$$GAS = 50 + \frac{10\Sigma(w_i x_i)}{\sqrt{(0.7\Sigma w_i^2 + 0.3(\Sigma w_i)^2)}}$$

w_i = the weight (*importance x difficulty*) assigned to the ith goal
x_i = the numerical value achieved for the ith goal

The same weights were used in calculating the score from both the therapist and the masked assessor. Consequently, differences in scores are all attributable to differences in the actual ratings of the goals.

Reliability was investigated using a mixed model intra-class correlation coefficient ($ICC_{(A,k)}$) (two-way mixed model with absolute agreement) (29). ICC values above 0.75 are considered to represent excellent reliability, values between 0.4 and 0.75 to represent fair to good reliability and values below 0.4 to represent poor reliability (30). The 95% limits of agreement (LoA = mean difference ± 1.96 standard deviation of the differences) (31, 32) were used to illustrate the agreement between the 2 scoring procedures. Normality of the data, absence of systematic bias and homoscedasticity were confirmed. Statistical Package for the Social Sciences (SPSS) software version 17.0 was used for analyses.

The actual goals set were categorized into groups based on the Rehabilitation Activities Profile (33).

RESULTS

Data from 29 patients (of 30 recruited) were used. Two patients had 3 goals each and one patient had 2 goals, giving a total of 112 goals. Table I presents some descriptive data of the 29 patients included in this study at baseline. Two patients could not complete the Short Orientation Memory Concentration Test (23). The mean (SD) GAS scores by the therapist and the assessor at 6 weeks are also presented.

Table II presents the goal areas covered in categories based on the Rehabilitation Activities Profile (33) with two additional domains specific to arm and leg activities that were not covered by the Rehabilitation Activities Profile but were evident from the goals. A wide variety of goals was used, with mobility and personal care being the largest domains.

Reliability and agreement

The mixed model $ICC_{(A,k)}$ between the therapist and the masked assessor scoring procedures is 0.478.

Fig. 1 shows a plot of the difference between the measurements (therapist–assessor) by the two procedures for each subject against their mean, including the Limits of Agreement (LoA) (–1.52 ± 24.54). Normal distribution of the differences and absence of systematic bias and heteroscedasticity were confirmed.

DISCUSSION

This study shows goal attainment scored by a treating therapist had low agreement with attainment scored by an independent assessor, although there was no systematic difference. If goal

Table I. *Descriptive data for the research population at baseline and the goal attainment scaling (GAS) score after 6 weeks*

Variable	Result (mean (SD))
Gender	11 females/18 male
Diagnosis	
Stroke	*n* = 27
Traumatic brain injury	*n* = 1
Multiple sclerosis	*n* = 1
Age	50.28 (13.88)
Time since onset (weeks) (*n* = 28)	18.86 (16.19)
Short Orientation Memory Concentration Test (*n* = 27)	22.22 (4.77)
Motricity Index (*n* = 29)	
UL	58.38 (31.38)
LL	56.00 (26.15)
Total	57.19 (25.45)
Barthel Index (*n* = 29)	12.17 (6.62)
Rivermead Mobility Index (*n* = 29)	6.38 (5.40)
NEADL (*n* = 29)	19.90 (14.86)
ARAT (*n* = 29)	25.59 (22.89)
GAS score (*n* = 29)	
Therapist	51.99 (11.01)
Assessor	53.51 (10.29)

The patient with multiple sclerosis was excluded from the calculation of the time since onset because this was an outlier (10 years).
SD: standard deviation; UL: upper limb; LL: lower limb; NEADL: Nottingham Extended ADL scale; ARAT: Action Research Arm Test.

Table II. *Goal areas according to the activities from the Rehabilitation Activities Profile plus two categories specific to upper and lower limb activities*

Category	Sub-category	Number of goals
Communication ($n=5$)	Expressing	5
Mobility ($n=42$)	Maintaining posture	4
	Changing posture	11
	Walking	20
	Using wheelchair	1
	Climbing stairs	6
Personal care ($n=38$)	Eating and drinking	17
	Washing and grooming	9
	Dressing	11
	Maintaining continence	1
Occupation (7)	Providing for meals	3
	Professional activities	1
	Leisure activities	3
Upper limb specific activities		18
Lower limb specific activities		2
Total		112

attainment scores are to be compared between patients or groups, more reliable scoring needs to be achieved.

In this study differences in training or skills between the independent and the treating assessors are unlikely. All were experienced in treating neurologically disabled patients and had similar training in the scaling and scoring procedures.

The most likely explanation for the different scoring lies in the method of obtaining the information needed to allocate a score. The independent assessor was masked to treatment and thus unfamiliar with the patient. Treating therapists were inevitably familiar with the actual abilities and performance of their patients, and could allocate a score on the basis of observation and interaction over the preceding few days.

Independent assessors inevitably had no prior information about the patient and had to extract it all in one session. Although some target activities could easily be observed, others could not because: (*i*) goals involved activities that could have compromised the safety of the patient and/or the assessor (e.g. climbing stairs or making a hot drink); (*ii*) goals required equipment not readily available (e.g. a kettle for boiling water); and (*iii*) goals involved observing behaviour in particular situations or settings (e.g. communicating with a partner using an alphabet chart). Thus the assessor depended upon verbal report, usually from the patient. Deficits in a patient's cognition and communication may thus have affected the scoring of the goals by the assessor. The best source of information, the treating therapist, was not available to the independent assessor.

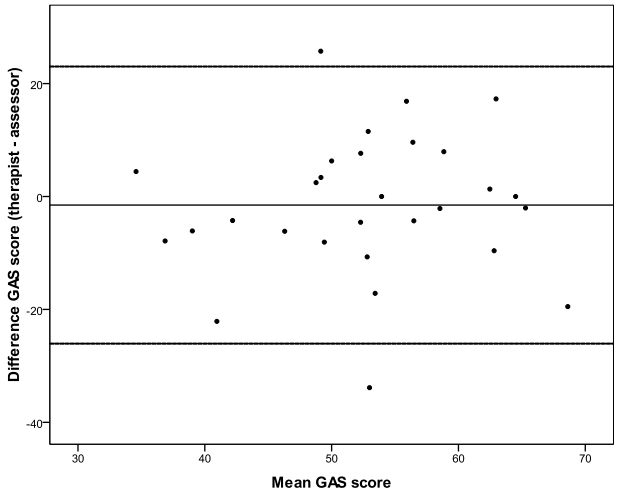

Fig. 1. Bland-Altman plot of the goal attainment scailing (GAS) score ($n=29$). Difference in GAS score between therapist and assessor against mean GAS score. Limits of agreement: (−1.52 ± 24.54).

In addition, some error may have arisen from ambiguity and uncertainty about the precise level of performance described. The therapist who sets a goal inevitably will retain additional information about the goal set, whereas the independent assessor only has the text. Thus there may have been some variation in interpretation of the descriptions determining the score. However, any variability in interpretation must have been both ways because there was no systematic bias favouring one class of assessor.

We do not have additional data allowing us to analyse this variability any further. It is not known whether two treating therapists or two independent assessors would vary as much. There is no data justifying the scoring decisions made by treating therapists. Individual classes of goal have not been analysed, not least because the numbers are small in many groups.

The potential advantages of goal attainment scaling as an outcome measure in patients with complex disabilities are its person-centred approach, the quantitative assessment of goal achievement, the lack of floor and ceiling effects and its responsiveness (21). However, there is also some controversy. Some authors challenge the mathematical concepts of goal attainment scaling, such as its non-linearity (34, 35) and the lack of uni-dimensionality (36), whereas others have raised concern about the validity of goal weighting (37). There is a lack of large-scale inter-rater reliability studies, and the actual scoring methods are usually described poorly and vary between studies (see below). Practically, goal attainment scaling can be unwieldy, time-consuming and requires knowledge and training for the clinicians.

Other studies have scored the goal attainment in different ways, such as through consensus within the clinical team (14, 17) or using a telephone interview method with the patient to score the goals (4). There is no evidence on whether these are more or less reliable. Inter-rater reliability of goal attainment scoring was previously reported to be good (12–17) but to our knowledge this is the first study investigating the agreement between two different goal attainment scoring procedures.

There are other methods for personalizing goals, such as the Canadian Occupational Performance Measure (38), but we are unaware of any research into the comparison between treating therapists and independent assessors with these measures.

The reliability observed in this study is poor compared with studies of the reliability of standardized measures such as the Barthel ADL index (27, 39, 40) or Rivermead Mobility Index (25, 27). Given that, in practice, the main goals set related to mobility and personal activities of daily living there is at least an argument that goal attainment scaling is not necessary in a population with neurological inpatients.

In conclusion, the attraction of goal attainment scaling as a sensitive and personalized measure of outcome suitable for use in randomized trial of complex interventions in heterogeneous groups of patients may be countered by the loss of investigational power arising from low reliability when measured by different procedures. Further studies are urgently needed. In the meantime, for inpatient populations, standardized measures may remain the best choice because existing measures cover the main areas of concern to patients.

ACKNOWLEDGEMENTS

The authors would like to thank the therapists at the Oxford Centre for Enablement (UK), Charlotte Winward and Emad El-Yahya, for their help in this study and Joan Warren for her financial support.

REFERENCES

1. Hurn J, Kneebone I, Cropley M. Goal setting as an outcome measure: a systematic review. Clin Rehabil 2006; 20: 756–772.
2. Levack WM, Taylor K, Siegert RJ, Dean SG, McPherson KM, Weatherall M. Is goal planning in rehabilitation effective? A systematic review. Clin Rehabil 2006; 20: 739–755.
3. Wade DT. Goal setting in rehabilitation: an overview of what, why and how. Clin Rehabil 2009; 23: 291–295.
4. Bouwens SF, van Heugten CM, Verhey FR. The practical use of goal attainment scaling for people with acquired brain injury who receive cognitive rehabilitation. Clin Rehabil 2009; 23: 310–320.
5. Bovend'Eerdt TJ, Dawes H, Sackley C, Izadi H, Wade DT. An integrated motor imagery program to improve functional task performance in neurorehabilitation: a single-blind randomized controlled trial. Arch Phys Med Rehabil 2010; 91: 939–946.
6. Brock K, Black S, Cotton S, Kennedy G, Wilson S, Sutton E. Goal achievement in the six months after inpatient rehabilitation for stroke. Disabil Rehabil 2009; 31: 880–886.
7. Hofer H, Holtforth MG, Frischknecht E, Znoj HJ. Fostering adjustment to acquired brain injury by psychotherapeutic interventions: a preliminary study. Appl Neuropsychol 2010; 17: 18–26.
8. Khan F, Pallant JF, Turner-Stokes L. Use of goal attainment scaling in inpatient rehabilitation for persons with multiple sclerosis. Arch Phys Med Rehabil 2008; 89: 652–659.
9. Turner-Stokes L, Baguley IJ, De Graaff S, Katrak P, Davies L, McCrory P, et al. Goal attainment scaling in the evaluation of treatment of upper limb spasticity with botulinum toxin: a secondary analysis from a double-blind placebo-controlled randomized clinical trial. J Rehabil Med 2010; 42: 81–89.
10. Turner-Stokes L, Williams H, Johnson J. Goal attainment scaling: does it provide added value as a person-centred measure for evaluation of outcome in neurorehabilitation following acquired brain injury? J Rehabil Med 2009; 41: 528–535.
11. Kiresuk TJ, Sherman RE. Goal attainment scaling: a general method for evaluating comprehensive community mental health programs. Community Ment Health J 1968; 4: 443–453.
12. Joyce BM, Rockwood KJ, Matekole CC. Use of goal attainment scaling in brain injury in a rehabilitation-hospital. Am J Phys Med Rehabil 1994; 73: 10–14.
13. Rockwood K, Joyce B, Stolee P. Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. J Clin Epidemiol 1997; 50: 581–588.
14. Rockwood K, Stolee P, Fox RA. Use of goal attainment scaling in measuring clinically important change in the frail elderly. J Clin Epidemiol 1993; 46: 1113–1118.
15. Rushton PW, Miller WC. Goal attainment scaling in the rehabilitation of patients with lower-extremity amputations: a pilot study. Arch Phys Med Rehabil 2002; 83: 771–775.
16. Stolee P, Rockwood K, Fox RA, Streiner DL. The use of goal attainment scaling in a geriatric care setting. J Am Geriatr Soc 1992; 40: 574–578.
17. Stolee P, Stadnyk K, Myers AM, Rockwood K. An individualized approach to outcome measurement in geriatric rehabilitation. J Gerontol A Biol Sci Med Sci 1999; 54: M641–M647.
18. Rockwood K, Howlett S, Stadnyk K, Carver D, Powell C, Stolee P. Responsiveness of goal attainment scaling in a randomized controlled trial of comprehensive geriatric assessment. J Clin Epidemiol 2003; 56: 736–743.
19. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006; 59: 1033–1039.
20. Bovend'eerdt TJ, Botell RE, Wade DT. Writing SMART rehabilitation goals and achieving goal attainment scaling: a practical guide. Clin Rehabil 2009; 23: 352–361.
21. Turner-Stokes L. Goal attainment scaling (GAS) in rehabilitation: a practical guide. Clin Rehabil 2009; 23: 362–370.
22. Blake H, McKinney M, Treece K, Lee E, Lincoln NB. An evaluation of screening measures for cognitive impairment after stroke. Age Ageing 2002; 31: 451–456.
23. Wade DT, Vergis E. The Short Orientation-Memory-Concentration Test: a study of its reliability and validity. Clin Rehabil 1999; 13: 164–170.
24. Collen FM, Wade DT, Bradshaw CM. Mobility after stroke: reliability of measures of impairment and disability. Int Disabil Stud 1990; 12: 6–9.
25. Forlander DA, Bohannon RW. Rivermead Mobility Index: a brief review of research to date. Clin Rehabil 1999; 13: 97–100.
26. Wade DT, Collin C. The Barthel ADL Index: a standard measure of physical disability? Int Disabil Stud 1988; 10: 64–67.
27. Green J, Forster A, Young J. A test-retest reliability study of the Barthel Index, the Rivermead Mobility Index, the Nottingham Extended Activities of Daily Living Scale and the Frenchay Activities Index in stroke patients. Disabil Rehabil 2001; 23: 670–676.
28. Platz T, Pinkowski C, van Wijck F, Kim IH, di Bella P, Johnson G. Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study. Clin Rehabil 2005; 19: 404–411.
29. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996; 1: 30–46.
30. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.
31. Bland JM, Altman DG. Statistical-methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–310.
32. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8: 135–160.
33. van Bennekom CA, Jelles F, Lankhorst GJ, Bouter LM. The Rehabilitation Activities Profile: a validation study of its use as a disability index with stroke patients. Arch Phys Med Rehabil 1995; 76: 501–507.
34. Steenbeek D, Ketelaar M, Galama K, Gorter JW. Goal attainment scaling in paediatric rehabilitation: a critical review of the literature. Dev Med Child Neurol 2007; 49: 550–556.
35. Steenbeek D, Meester-Delver A, Becher JG, Lankhorst GJ. The effect of botulinum toxin type A treatment of the lower extremity on the level of functional abilities in children with cerebral palsy: evaluation with goal attainment scaling. Clin Rehabil 2005; 19: 274–282.
36. Tennant A. Goal attainment scaling: current methodological challenges. Disabil Rehabil 2007; 29: 1583–1588.
37. Mackay G, Somerville W, Lundie J. Reflections on goal attainment scaling (GAS): cautionary notes and proposals for development. Educ Res 1996; 38: 161–172.
38. Carswell A, McColl MA, Baptiste S, Law M, Polatajko H, Pollock N. The Canadian Occupational Performance Measure: a research and clinical literature review. Can J Occup Ther 2004; 71: 210–222.
39. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. Int Disabil Stud 1988; 10: 61–63.
40. Gosman-Hedstrom G, Svensson E. Parallel reliability of the functional independence measure and the Barthel ADL index. Disabil Rehabil 2000; 22: 702–715.