ORIGINAL ARTICLE

Interrater Reliability of Goal Attainment Scaling in Rehabilitation of Children With Cerebral Palsy

Duco Steenbeek, MD, Marjolijn Ketelaar, PhD, Eline Lindeman, MD, PhD, Krys Galama, OT, Jan Willem Gorter, MD, PhD

ABSTRACT. Steenbeek D, Ketelaar M, Lindeman E, Galama K, Gorter JW. Interrater reliability of goal attainment scaling in rehabilitation of children with cerebral palsy. Arch Phys Med Rehabil 2010; 91:429-35.

Objectives: To determine the interrater reliability of Goal Attainment Scaling (GAS) in the routine practice of interdisciplinary rehabilitation of children with cerebral palsy, and to examine the difference in the interrater reliability of the scores between GAS scales constructed by the children's own therapists and the scales constructed by independent therapists.

Design: Individually tailored GAS scales, based on predetermined criteria, were constructed at the start of a 6-month rehabilitation period. The outcome was rated independently by 2 therapists at the end of the treatment period. Two different data sets were acquired, one consisting of scores on GAS scales constructed by the children's own therapists, the other of scores on GAS scales constructed by matched independent raters of the same profession.

Setting: A children's unit of a medium-sized rehabilitation center in The Netherlands.

Participants: Physical therapists (n=8), occupational therapists (n=8), and speech therapists (n=4) participated in pairs. They constructed 2 sets of 64 GAS scales each, for 23 children with cerebral palsy.

Interventions: A 6-month interdisciplinary pediatric rehabilitation program.

Main Outcome Measure: Interrater reliability was assessed using linear-weighted Cohen's kappa.

Results: The scales constructed by the children's therapists had an interrater reliability of .82 (95% confidence interval [CI], .73–.91). The interrater reliability for scales constructed by the independent raters was .64 (95% CI, .49–.79). The main reason for disagreement between raters was discrepancies in the professionals' interpretation of the children's capacities versus their actual performance during assessment.

Conclusions: The interrater reliability of GAS used under optimal conditions was good, particularly for scales constructed by the children's own therapists.

Key Words: Cerebral palsy; Disability evaluation; Goals; Outcome assessment (health care); Rehabilitation; Reproducibility of results.

AN IMPORTANT ASPECT of pediatric rehabilitation is to evaluate patients' and families' progress towards activity and participation goals. Given the diversity of developmental disabilities and therapy goals, this requires the use of individual measurement tools. GAS is an increasingly popular individual instrument for progress assessment in rehabilitation. Kiresuk and Sherman[1] introduced GAS to evaluate adult mental health services. In its original form, it is a 5-point scale, constructed before an intervention period, with "0" representing the expected level of functioning after a predefined period. If a patient achieves more than is expected, a score of +1 or +2 is given, depending on the level of achievement. If the patient's progress is less than expected, a score of −1 or −2 is given. Several recent studies have experimented with scales using 6,[2] 7,[3] or 3[4] points, and in most research with GAS, "no change" is scored as −2. In a previous article, we presented arguments in favor of using a score of −3 for deterioration.[5] In this version of GAS, the additional score of −3 allows all original levels to be attained and avoids bottoming effects. This removes one of the causes of false sensitivity to change in calculating group effects.[5]

The sensitivity of GAS to change is assumed to be better than that of standardized functional measures,[5,6] while the reliability of the method of scale development (content reliability) and the reliability of the scores (interrater reliability) in predetermined scales are still subject to debate. Despite this, GAS has been used extensively in recent intervention studies—for example, to assess botulinum toxin A treatment in the pediatric population.[7-9] Research into adult mental health and geriatric medicine has provided evidence about the psychometric properties of GAS, reporting the content reliability of GAS to be good and its interrater reliability satisfactory.[10-12]

Knowledge about the reliability of GAS in rehabilitation care is limited, and further validation and reliability studies to justify the use of GAS are warranted.[5,13] Palisano[14] examined the interrater reliability of GAS in pediatric physical therapy in the context of a validation study. He reported good interrater reliability, with kappa coefficients of .89 for 10 scales scored

List of Abbreviations

CI	confidence interval
CP	cerebral palsy
GAS	Goal Attainment Scaling
GMFCS	Gross Motor Function Classification System

From the Departments of Physical Medicine and Rehabilitation (Steenbeek) and Occupational Therapy (Galama), Rehabilitation Center Breda, Breda, The Netherlands; the Departments of Research and Development, Rehabilitation Center De Hoogstraat, Utrecht, The Netherlands (Ketelaar, Lindeman); the Department of Rehabilitation and Sports Medicine, University Medical Center, and Rudolf Magnus Institute of Neuroscience, Utrecht, The Netherlands (Ketelaar, Lindeman); Partner in NetChild, Network for Childhood Disability Research in the Netherlands, Utrecht (Steenbeek, Ketelaar, Gorter); and CanChild Center for Childhood Disability Research, McMaster University, Hamilton, Ontario, Canada (Gorter).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit on the authors or on any organization with which the authors are associated.

Reprint requests to Duco Steenbeek, MD, Rehabilitation Center Breda, Brabantlaan 1, 4817 JW Breda, The Netherlands, e-mail: *d.steenbeek@rcbreda.nl.*

0003-9993/10/9103-00473$36.00/0
doi:10.1016/j.apmr.2009.10.013

before the start of a validation study, and .75 for 16 scales scored during the validation study. In a previous study by Palisano et al,[15] a physical therapist and an occupational therapist independently scored the performance of 9 subjects on a GAS scale, and their scores all matched. These authors constructed GAS scales using information from the therapists, and then scored them simultaneously with an independent rater from a video. The sample sizes in both studies were, however, small. Despite Palisano's recommendation to replicate interrater reliability studies across therapists and clients of various ages and disabilities, no further studies have been performed in this field in the last 15 years.

As shown in a previous study,[16] the reliability of the scores largely depends on the agreements made about scale development and scoring procedures, and on experience with GAS. Further studies showing that GAS has acceptable interrater reliability in practical settings are essential if this method is to be used in rehabilitation practice.

The reliability of GAS may be affected by the risk of so-called therapist bias. One potential source of therapist bias is that of therapists' expectations about their patients' level of attainment for each GAS scale. Another source of bias arises when therapists score their own scales, given the probable dependence of the therapists and their interest in a good outcome.[15] To decrease the risk of therapist bias, stringent adherence to the original GAS protocol dictates that goal setters and raters be independent of the treatment process.[17] However, the influence of goal setters and/or raters being involved in the treatment process has never been examined, and independent construction and scoring procedures for GAS are, in fact, impractical and inefficient in routine rehabilitation. The functional relevance of independent GAS construction is debatable, as goals should be set together with the children and their families and should take children's changeability into account. The therapists are the people who have optimal knowledge of their patients' history and treatment results, and this knowledge is essential for establishing the actual goals and different GAS scale levels.

The aims of this study were (1) to determine the interrater reliability of GAS in the routine practice of children's therapists working in a team of trained professionals in interdisciplinary pediatric rehabilitation and explore the reasons for discrepancies between scores; and (2) to examine the difference in the interrater reliability of the scores between GAS scales constructed by the children's own therapists and that of scores on scales constructed by independent therapists.

METHODS

Design

A convenience sample of 20 members of a trained team of professionals, consisting of physical, occupational, and speech therapists, participated in the reliability study with a pretest-posttest design. Two team members of each of the 3 disciplines were selected as independent raters. Fixed pairs were formed consisting of each primary child's therapist and one of the independent raters. If one of the selected independent raters worked as the child's therapist, that rater filled the role of the child's therapist, and the second independent rater of that discipline was paired with him/her. Demographic baseline data on the raters (age and years of professional experience) were recorded.

The cases included were children with CP of various degrees of severity. Children were eligible for this study if they were between the ages of 2 and 14 years and their physician expected them to be in interdisciplinary therapy for at least 6 months. The distribution of severity of CP was evaluated by the GMFCS[18] and the Manual Ability Classification System.[19]

The study leader constructed a case summary for each child, based on the request for help and the expectations of each child and the child's parents. Each case summary included separate domain descriptions for the different disciplines. The child's therapist and the independent rater constructed GAS scales individually, both having been asked to base their scale development on the study leader's case summary and interpretation of the child's functioning. The child's therapists constructed GAS during regular therapy, whereas the independent raters constructed GAS during a separate 1-hour observation of the child, who was accompanied by his/her parents. Goals were set for a 6-month period. After all GAS scales for a child had been constructed, the scales were shared with the primary therapists and the child's parents. After the 6 months, the child's therapist and the independent rater both scored the GAS scale constructed by the child's therapist, as well as the GAS scale constructed by the independent rater (table 1). The child's therapists scored GAS during regular therapy, whereas the independent raters scored the scales in a separate 30-minute session after the therapy period.

To guarantee independence between the members of each pair, the child's therapist and independent rater were asked to avoid exchanging any details about the child's progress during therapy. During a meeting with the parents, the study leader explicitly asked the parents not to discuss the construction of

Table 1: Design of the Reliability Study and Actual Number of GAS Scales per Discipline

	Scored by the Child's Therapist			Scored by the Independent Rater		
GAS Scales	PT	OT	ST	PT	OT	ST
Constructed by the child's therapist						
PT	23	ND	ND	23	ND	ND
OT	ND	23	ND	ND	23	ND
ST	ND	ND	18	ND	ND	18
Totals A		64			64	
Constructed by the independent rater						
PT	23	ND	ND	23	ND	ND
OT	ND	23	ND	ND	23	ND
ST	ND	ND	18	ND	ND	18
Totals B		64			64	

NOTE. The first aim of this study was to determine the interrater reliability within the GAS scales constructed by the child's therapist (ie, the totals of A). The second aim was to compare the reliability data for the GAS scales constructed by the child's therapists (ie, the totals of A) with those for the GAS scales constructed by the independent raters (ie, the totals of B).
Abbreviations: ND, no data; OT, occupational therapy; PT, physical therapy; ST, speech therapy.

the scales or the scoring, to help ensure the independence between the child's therapist and independent rater.

Goal Attainment Scaling Method Used

Before the start of the present study, an 8-month practical training period was used to introduce GAS in the children's rehabilitation unit of a medium-sized rehabilitation center in The Netherlands. In the training program, which we described and evaluated in detail in a previous article,[16] professionals practiced the construction and scoring of GAS scales. Six-point GAS scales were constructed, using −3 for deterioration and −2 for no change relative to the level determined at the start of treatment, as described in Steenbeek et al.[16] The participants agreed to adhere to the following criteria for scale development. All goals should be based on the request for help in the area of a child's capacity for daily activities. The goals must be important for and relevant to the children and their families, and describe the main aim of therapy for each discipline involved. Each level should be defined as clearly as possible. All levels of the scales should be specific, measurable, acceptable, realistic, and time-specific. The therapists were asked to construct ordinal scales with incremental steps of equal intervals. Each GAS scale had to reflect a single dimension of change. In addition, it had to be possible to score a scale within 10 minutes to ensure that it was practicable. The setting and task were described explicitly for each GAS scale.

Scale rating. All raters were asked to base the scores on their insight as a professional of goal achievement rather than on the child's actual performance at the time of the assessment. The reflection of the professionals' knowledge and experience in rating goal attainment was introduced to minimize the possible influence of children's whims (fatigue, motivation, interaction with the therapist, and behavioral issues). The raters' professional judgment of goal attainment was supported by observing the child or by conferring with parents or teachers, and the method was stipulated in each GAS scale description. When conferring was stipulated, the raters informally asked the parents or teachers their opinion about the child's functioning.

Evaluation of Goal Attainment Scaling Scores

To meet the first aim of the study, we compared the scores given by the pairs on the GAS scales constructed by the children's own therapists, to evaluate the interrater reliability. For the second aim, the scores given by the pairs on GAS scales constructed by the independent raters were used as a second data set. The interrater reliability found for the 2 data sets were then compared to determine the influence of independent scale construction on the interrater reliability of the scores. The design was such that the 2 raters rating the same GAS scale had different perspectives, for example, in terms of their knowledge of the child's history. An additional evaluation of the scores was performed when the raters of a pair disagreed, for a better understanding of the consequences of independent scale construction.

After scoring had been completed, therapists were informed about the discrepancies in their scoring, thus ending the independence within the pairs. The pairs were then systematically interviewed to explore reasons for the scoring discrepancies.

Further aspects of GAS construction and scoring that might have influenced the interrater reliability were explored. We calculated the interrater reliability per discipline to gain insight in the separate CIs, as scores from different disciplines for the same child could be dependent. The influence of the type of GAS scale on the interrater reliability was also determined. After scoring was complete, the study leader classified all GAS scales into 2 categories, one with scales rating physical function and the other rating higher cognitive function. Scales rating physical function, such as running distance, are more concrete and observable, whereas scales rating higher cognitive function, such as recognizing clothes when self-dressing or self-efficacy in maneuvering a wheelchair, depend more on interpretation. This subdivision was checked by one of the authors (KG). Interrater reliability was then calculated for each type of GAS scale.

Data Analysis

The interrater reliability was evaluated by calculating linear-weighted Cohen's kappa values with 95% CIs. The kappa values were interpreted as no (<0), very low (0.0–.20), low (.21–.40), moderate (.41–.60), good (.61–.80), or excellent agreement (.81–1.00).[20] We used the Wilcoxon signed-rank test for each data set to test whether the scores were more likely to be higher or lower when rated by the child's own therapists than when rated by the independent raters. Statistics were performed in SPSS/PC, release 14.0[a] and http://faculty.vassar.edu/lowry/kappa.html.

RESULTS

Eight physical therapists, 8 occupational therapists, and 4 speech therapists participated. The 20 professionals involved (age range, 28–59y; mean age ± SD, 40.3±10.9y) had 4 to 30

Table 2: Summary of Characteristics of 23 Children With CP

Patient	Age (y)	Sex	GMFCS Level	MACS Level	No. of Disciplines*
1	12	B	IV	IV	3
2	13	G	V	V	2
3	7	G	II	I	3
4	13	G	V	V	3
5	5	B	I	III	3
6	7	B	III	II	3
7	9	G	IV	II	3
8	11	G	III	II	3
9	4	B	I	III	3
10	7	B	II	I	3
11	6	B	V	V	3
12	3	B	II	III	3
13	5	B	III	II	2
14	8	B	IV	IV	3
15	9	G	II	I	2
16	10	B	III	II	2
17	6	B	II	II	3
18	6	B	I	III	3
19	10	G	V	V	3
20	10	B	IV	II	3
21	2	B	II	III	3
22	11	G	V	V	3
23	5	B	II	I	2

NOTE. GMFCS[18] classifies the motor function of children with CP based on their self-initiated movement with particular emphasis on sitting, walking, and wheeled mobility. Level I represents walking without limitations at 6 to 12 years. Level V represents no means of independent mobility at that age (http://www.canchild.ca). MACS[19] classifies how children with CP use their hands when handling objects as part of daily activities. Level I represents the best functional abilities and level V the most limited ones (http://www.macs.nu).

Abbreviations: B, boy; G, girl; MACS, Manual Ability Classification System.

*Number of disciplines involved: 3, physical therapy, occupational therapy, and speech therapy; 2, physical therapy and occupational therapy only.

Table 3: Example of a GAS Scale*

Case summary: L is a 5-year-old boy with right-sided unilateral spastic CP, GMFCS level II. The main request for help is to fall less often. L falls a few times a day while standing on his right leg during normal walking. Treatment: a multilevel botulinum toxin intervention and physical therapy 4 times a week for 4 months.

	Child's Physical Therapist	Physical Therapist Independent Rater
Setting	A strictly specified gymnasium with an obstacle course including jumping and quick changes of walking direction. Safety is guaranteed by guidance. The therapist encourages L to complete the course within 3min.	L is wearing his ankle-foot orthosis and shoes. Observation.
Scoring Method	Observation	Observation
Task	Walk the obstacle course fast and don't fall.	Stand on your right leg only as long as possible.
−3 Worse than start (deterioration)	L falls 5 times or more.	<2s
−2 Equal to start	L falls 4 times.	2s
−1 Less than expected	L falls 3 times.	3–5s
0 Expected goal	L falls 2 times.	6–8s
1 Somewhat more than expected	L falls 1 time.	9–11s
2 Much more than expected	L does not fall.	>11s

*The GAS scale was constructed at the start of an intervention period by a child's own physical therapist and the physical therapist functioning as independent rater, based on the case summary provided by the study leader (No. 23 in table 2). After 6 months of rehabilitation, the first resulted in agreement on the score, the second in disagreement.

years (mean ± SD, 14.8±9.8y) of professional experience. Before the training, none of the therapists or independent raters had had any experience with GAS.[16]

GAS scales were constructed for 23 children in the reliability study. Patients ranged in age from 4 to 13 years, and severity of CP ranged from GMFCS level I to GMFCS level V (table 2). Therapists treated 1 to 3 children each. In 18 cases, one of the therapists appointed as independent rater was the child's primary therapist.

The total number of GAS scales constructed was 18 × 6 (3 therapists + 3 independent raters) plus 5 × 4 (2 therapists + 2 independent raters) for a total of 128 scales (see table 1). Table 3 shows an example of a case summary, with the domain description for the physical therapists, and the GAS scales constructed by a physical therapist and his matched independent rater. The different approaches and descriptions by the raters forming the pair illustrate their independence.

Table 4 is the cross table of the scores given by therapists and independent raters on the GAS scales constructed by the children's own therapists after 6 months of rehabilitation (n=64), showing 77% agreement between the raters forming the pairs and a linear-weighted Cohen's kappa of .82 (95% CI, .73–.91), indicating excellent agreement.

Table 5 is the cross table of the scores given by therapists and independent raters on GAS scales constructed by the independent raters after 6 months of rehabilitation (n=64), showing 64% agreement between raters forming the pairs and a linear-weighted Cohen's kappa of .64 (95% CI, .49–.79), indicating good agreement.

On the GAS scales constructed by the children's own therapists, the scores given by the 2 groups of raters matched for 49 scales, while the children's therapists gave higher scores on 6 scales, and the independent raters gave higher scores on 9 scales (see table 4). Comparison of these scores revealed no statistically significant difference in scoring patterns (Wilcoxon signed-rank test, $z=-.62$; $P=.54$). Among the scales constructed by the independent raters, there were 11 on which the children's therapists gave a higher score, and 12 where the independent raters gave a higher score (see table 5). There was no statistically significant difference in these scoring patterns either ($z=-.16$; $P=.88$). There were 15 discrepancies between raters scoring the scales constructed by the children's own therapists (2 × 2 points; 13 × 1 point [see table 4]) and 23 between raters scoring the scales constructed by the independent raters (9 × >1 point; 14 × 1 point [see table 5]).

Systematic interviews to evaluate discrepancies between the pairs in both data sets revealed that disagreement was largely due to different interpretations of the professionals' perception of the child's capacity, as opposed to what a child actually does when tested. In 12 scales, one rater unintentionally scored the actually observed performance, although he realized the child did not perform to the best of his ability. Differences in 10 scales were caused by a difference of opinion between the raters, when children were capable of more than they actually did during the assessment. Differences in 4 other scales were due to one rater misunderstanding the textual content of the scale (eg, a standard obstacle course in a gymnasium was not specific enough for the partner, who set up a more complex course). Causes of 12 discrepancies remained unclear.

The kappa values for each discipline are shown in table 6. The mean kappa value corresponded with that of the 2 entire data sets. The kappa values for each type of scale are shown in table 7. The study leader (DS) and another author (KG) agreed on the subdivision of the types of scales for 95% of the scales, and after discussion agreed on 100% of them. Thirty percent of the physical scales resulted in disagreement between the pairs of raters versus 29% of the higher cognitive function scales.

DISCUSSION

This study found good-to-excellent interrater reliability of GAS when used by a group of trained therapists for children with CP. Similar to the studies on the reliability of GAS by Palisano,[14,15] this study provides further evidence for the interrater reliability of GAS in a regular rehabilitation setting.

Table 4: Scores on the GAS Scales Constructed by the Children's Own Therapists to Calculate the Interrater Reliability*

GAS Score by the Independent Rater	GAS Score by the Child's Therapist						
Score	−3	−2	−1	0	1	2	Totals
−3	0	0	0	0	0	0	0
−2	0	5	0	0	0	0	5
−1	0	4	12	3	0	1	20
0	0	0	0	8	0	0	8
1	0	0	1	1	10	2	14
2	0	0	0	0	3	14	17
Totals	0	9	13	12	13	17	64

*Linear-weighted Cohen's kappa was .82 (95% CI, .73–.91).

This is reassuring because GAS has been used extensively in recent rehabilitation research.

The present study confirms the necessity of further discussion of GAS training procedures, as previously recommended.[16] The reliability of the scores depends on the agreements made regarding scale development and scoring procedures, as well as the experience gained with these procedures. Because there were still misunderstandings between the members of the pairs of raters, even after training, the reliability of the GAS would most likely have been lower without training. The training period in this study was intended to improve the homogeneity of the method performed by all raters, as the interrater reliability was assessed between groups of raters with divergent levels of professional experience.

Critics of GAS have doubted the value of the patient's own therapists selecting goals and specifying outcomes. Cytrynbaum et al[17] recommended that goal setters and raters be independent of the treatment process. Therefore, the second aim of this study was to examine the difference in interrater reliability of the scores on GAS scales constructed by the children's own therapists and those on scales constructed by therapists functioning as independent raters. The differences found between the groups (κ, .82 vs .64) were not statistically significant because of an overlap in the 95% CIs. The difference can be explained by either coincidence or the difference in the content of the scales constructed by the 2 groups. Familiarity with a child's history, character, and preferences, as well as treatment results, is probably needed to construct GAS scales that can be rated reliably. These data are reassuring as regards the presumed therapist bias in constructing GAS scales, and contradict Cytrynbaum's advice.

No significant differences in kappa values were found between the various disciplines or between the types of scales. Contrary to expectation, the scales rating physical function did not show a higher percentage of agreement. The good interrater reliability for scales rating higher cognitive function could be because we measured the professionals' insight rather than the child's performance during the assessment. The main reason for disagreement between raters was the interpretation of the children's capacity when they did not perform to the best of their ability during the assessment. Using the professional's judgment of the therapists and independent raters for scoring was assumed to be beneficial to the interrater reliability, although misinterpretation of this criterion can also be detrimental to the reliability. This emphasizes the importance of clear instructions.

Study Limitations

In terms of the first aim of our study, the results must be interpreted with caution, because the raters forming the pairs

Table 5: Scores on the GAS Scales Constructed by the Independent Raters to Compare the Interrater Reliability With That of Table 4*

GAS Score by the Independent Rater	GAS Score by the Child's Therapist						
Score	−3	−2	−1	0	1	2	Totals
−3	2	0	0	0	0	0	2
−2	0	5	1	0	0	1	7
−1	0	1	12	3	0	2	18
0	0	1	2	7	2	1	13
1	0	1	1	0	5	1	8
2	0	0	1	1	4	10	16
Totals	2	8	17	11	11	15	64

*Linear-weighted Cohen's kappa was .64 (95% CI, .49–.79).

Table 6: Kappa Values (95% CIs) per Discipline

Constructed by	Scales (n)	Scored by Pairs of PTs	Scales (n)	Scored by Pairs of OTs	Scales (n)	Scored by Pairs of STs
Child's therapists	23	.73 (.55–.90)	23	.84 (.68–1.00)	18	.92 (.82–1.00)
Independent raters	23	.66 (.43–.87)	23	.61 (.35–.86)	18	.65 (.34–.97)

Abbreviations: OTs, occupational therapists; PTs, physical therapists; STs, speech therapists.

differed in their knowledge of the child's history and treatment results. Because GAS intends to measure a professional's perception of a child's level of functioning, the children's own therapists had more information they could use in scoring than did the independent raters. This difference may have biased the interrater reliability. However, if this were true, the interrater reliability would have been underestimated rather than overestimated. No systematic differences in scoring patterns were found between therapists and independent raters.

Another limitation in regard to the first aim is that a possible dependence of the data could present a problem with the CIs of the kappa value, because the calculation was based on 64 GAS scales constructed for 23 children. However, 3 therapists, of different disciplines, each constructed a GAS scale in a different, predefined domain that was distinctive for their discipline. Three different independent pairs of therapists only scored the GAS scale constructed by their own discipline, to determine the agreement within the pairs. Although the 6 scores per child are dependent, this procedure makes it unlikely that there was appreciable dependence between the 3 values of agreement per child, confirming the 95% CIs of the kappa values given in the Results section. The kappa values could be biased by floor and ceiling effects. If serious deterioration (−3) or excellent performance (+2) in all domains of functioning caused dependence, the correlations could be overestimated. We had no floor effects. Ceiling effects were improbable because 19 of all 24 scales scored as +2 by both raters were for different children. Only the influence on the corresponding CIs has to be taken into consideration, because the mean kappa value of the separate data sets corresponds with the kappa of the entire data set. Multiple GAS scales could be constructed for a child, as usually more than one goal is set per child. GAS was originally characterized by the use of the T-sum formula, a mathematical technique quantifying the achievement in several weighted goals per therapist. Although this formula theoretically could correct for dependence, the correction is too subjective to provide a real solution for the dependence issue. The formula also introduces false sensitivity to change, as GAS data are at best ordinal.[21] Therefore we prefer to evaluate individual GAS scores and analyze multiple raw change scores with nonparametric statistics.[2]

A limitation of the study regarding the second aim is that the study may be difficult to reproduce, because of the case summary provided by the study leader and the instructions influencing communication between parents and therapists. Although the information in the case summary was kept to a minimum, it was subjective and may have influenced the independence during GAS scale construction between the therapists and independent raters. This subjectivity was, however, unavoidable, as the GAS scales of both raters had to be based on the child's and parents' request for help and their expectations. The influence of the children themselves on the independence between the raters is unknown.

A second limitation regarding the second aim is that the therapy theoretically focused more on the goals set by the child's therapists than on those set by the independent raters, enabling more accurate rating of the goals set by the child's therapists. However, after construction of both scales was completed, the independent rater's goal was shared with the child's therapist, and no instruction was given as to whether therapy should focus on GAS. Moreover, the outcomes for both groups of GAS scales were equal, with a median of 0.

Finally, the procedure used to measure goal attainment could have influenced the interrater reliability. That parents or teachers could be consulted as part of the judgment of goal attainment may have meant that the professionals' knowledge and experience had less influence on the score. The independence of the raters could decrease as a result of them both conferring with the same responder. We divided all GAS scales into 2 categories: scales where the score was supported by observing the child, and scales where the score was supported by conferring with parents or teachers, as was stipulated in each GAS scale. Thirty-two percent of the scales scored after observation resulted in disagreement, compared with 18% of the scales scored after consulting with parents or teachers.

At the end of the study, all the parents were interviewed and asked about their experience, in order to gain insight into the impact of participating in a study like this for parents and children. The parents had an active role in keeping the pairs of raters independent during the study, and the GAS scales were shared with the parents right after their construction. The parents of all 23 children indicated that they were satisfied with the use of GAS in the treatment of their child and with their participation in this study. Future reliability studies on GAS in other rehabilitation settings or with other diagnostic groups should be considered, as the burden on participating families is minimal.

The measurement properties of the content of GAS scales and the sensitivity to change are subjects that remain to be explored.[5] A possibility for future research is to have a team of experts rate the overall value of therapists' goals. The present study showed that the GAS scales constructed by the child's own therapists in particular should be used for further validity research. Even if an independent assessor were essential to reduce bias in evaluating outcome, the involvement of the child's own therapist (and the child or child's parents) is still necessary when designing goals for therapy.

Despite the need for future research, GAS is a very promising tool that is allowing children and their families to set transparent and important functional goals together with their health care professionals.

Table 7: Kappa Values (95% CIs) per Type of GAS Scales

Constructed by	Scales (n)	Physical	Scales (n)	Higher Cognitive Function
Child's therapists	21	.76 (.57–.95)	43	.85 (.75–.94)
Independent raters	18	.65 (.38–.92)	46	.63 (.46–.81)
All	39	.71 (.54–.87)	89	.74 (.63–.84)

CONCLUSIONS

In this study, GAS was used for children with CP, under optimal conditions (trained team using predetermined criteria) in the routine practice of a children's unit in a medium-sized rehabilitation center. The interrater reliability of GAS scores proved good to excellent. The results also suggest that scale construction by the child's own therapist as opposed to an independent rater has a positive influence on the interrater reliability of the scales. The interrater reliability can be further improved by standardizing the procedure. We found that discrepancies between the professionals' interpretation of the child's capacities and the child's actual performance during assessment were the main cause of disagreement between raters. Reliability studies in other rehabilitation settings and diagnostic groups are recommended. Further investigation of the content reliability and content validity in the construction of GAS scales and their sensitivity to change is also necessary and of great importance for rehabilitation.

Acknowledgments. We thank all participating children, parents, and therapists at the children's department of the Rehabilitation Center Breda, The Netherlands; Riekie H.C.W. de Vet, PhD, for her support as an expert on clinimetrics; and Tjeerd van der Ploeg for his statistical support.

References

1. Kiresuk T, Sherman R. Goal attainment scaling: a general method of evaluating comprehensive community mental health programs. Community Ment Health J 1968;4:443-53.
2. 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: 274-82.
3. Cusick A, McIntyre S, Novak I, Lannin N, Lowe K. A comparison of goal attainment scaling and the Canadian Occupational Performance Measure for paediatric rehabilitation research. Pediatr Rehabil 2006;9:149-57.
4. Ahl LE, Johansson E, Granat T, Carlberg EB. Functional therapy for children with cerebral palsy: an ecological approach. Dev Med Child Neurol 2005;47:613-9.
5. 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;49:550-6.
6. Mailloux Z, May-Benson TA, Summers CA, et al. Goal attainment scaling as a measure of meaningful outcomes for children with sensory integration disorders. Am J Occup Ther 2007;61: 254-9.
7. Lowe K, Novak I, Cusick A. Low-dose/high-concentration localized botulinum toxin A improves upper limb movement and function in children with hemiplegic cerebral palsy. Dev Med Child Neurol 2006;48:170-5.
8. Paolicelli PB. Use of botulinum toxin type A in walking disorders of children with cerebral palsy. Eur Med Phys 2001;37:83-92.
9. Mall V, Heinen F, Siebel A, et al. Treatment of adductor spasticity with BTX-A in children with CP: a randomized, double-blind, placebo-controlled study. Dev Med Child Neurol 2006;48:10-3.
10. Sheffler G, Canetti L, Wiseman H. Psychometric properties of goal-attainment scaling in the assessment of Mann's time-limited psychotherapy. J Clin Psychol 2001;57:971-9.
11. Rockwood K, Stadnyk K, Carver D, et al. A clinimetric evaluation of specialized geriatric care for rural dwelling, frail older people. J Am Geriatr Soc 2000;48:1080-5.
12. 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;56:736-43.
13. Schlosser RW. Goal attainment scaling as a clinical measurement technique in communication disorders: a critical review. J Commun Disord 2004;37:217-39.
14. Palisano RJ. Validity of goal attainment scaling in infants with motor delays. Phys Ther 1993;73:651-60.
15. Palisano RJ, Haley SM, Brown DA. Goal attainment scaling as a measure of change in infants with motor delays. Phys Ther 1992; 72:432-7.
16. Steenbeek D, Ketelaar M, Galama K, Gorter JW. Goal Attainment Scaling in paediatric rehabilitation: a report on the clinical training of an interdisciplinary team. Child Care Health Dev 2008;34: 521-9.
17. Cytrynbaum S, Ginath Y, Birdwell J, Brandt L. Goal attainment scaling: a critical review. Eval Q 1979;3:5-40.
18. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol 1997;39:214-23.
19. Eliasson AC, Krumlinde-Sundholm L, Rösblad B, et al. The Manual Ability Classification System (MACS) for children with cerebral palsy: scale development and evidence of validity and reliability. Dev Med Child Neurol 2006;48:549-54.
20. Landis JR, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
21. Tennant A. Goal attainment scaling: current methodological challenges. Disabil Rehabil 2007;15;1583-8.

Supplier

a. SPSS Inc, 233 S Wacker Dr, 11th Fl, Chicago, IL 60606.