Proposal of criteria for appraising Goal Attainment Scales used as outcome measures in rehabilitation research

Agata Krasny-Pacini, Jonathan Evans, Mckay Moore Sohlberg, Mathilde Chevignard

To cite this version:

HAL Id: hal-01203473
https://hal.sorbonne-universite.fr/hal-01203473
Submitted on 23 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Running head of no more than 40 character spaces: GAS in rehabilitation research

Title: Proposal of criteria for appraising Goal Attainment Scales used as outcome measures in rehabilitation research

Author(s) full name(s) written as First Name then Last Name, and highest academic degree(s):
Agata Krasny-Pacini¹,⁵,⁶, MD
Jonathan Evans², PhD
McKay Moore Sohlberg³, PhD
Mathilde Chevignard⁴,⁵, PhD

The name(s) of the institution(s), section(s), division(s), and department(s) where the study was performed and the institutional affiliation(s) of the author(s) at the time of the study:

1. Institut Universitaire de Réadaptation Clemenceau-Strasbourg, 45 bd Clemenceau, 67082 Strasbourg, France Paediatric Department and Hautepierre Strasbourg University Hospital, Strasbourg, France

2. Department of Psychological Medicine, Mental Health and Wellbeing, Gartnavel Royal Hospital, 1055 Great Western Rd, Glasgow G12 0XH, Scotland, UK

3. Communication Disorders & Sciences University of Oregon Eugene Oregon 97403 United States
4. Rehabilitation Department for Children with Acquired Brain Injury, Hôpitaux de Saint Maurice, 14, rue du Val d’Osne, 94410 Saint Maurice

5. Sorbonne Universités, UPMC Paris 06, UMR 7371, UMR_S 1146, LIB, F-75005, Paris, France

6. Service de Chirurgie Orthopédique Infantile, hôpital de Hautepierre, CHU de Strasbourg, avenue Molière, 67098 Strasbourg cedex

Acknowledgment of any presentation of this material, to whom, when, and where:
This work has not been presented yet.

Acknowledgment of financial support, including grant numbers: Writing this paper was partially supported by SOFMER (French society of PMR)-Allergan prize, won by the clinical team of Dr. Krasny-Pacini in October 2012.

We would like to thank Dr. Roumenoff-Turcant from Lyon for her feedback on the paper.

Explanation of any conflicts of interest: The authors report no conflict of interest

Name, address, business telephone number, and e-mail address of corresponding author:
Dr. Agata Krasny-Pacini (corresponding author)
Institut Universitaire de Réadaptation Clemenceau-Strasbourg
45 bd Clemenceau
67082 Strasbourg, France
Telephone: 0033388211642/47
0033 671-28-41-52 (mobile)
Fax : 0033 388211643/1605

E-mail: agatakrasny@yahoo.com
agata.krasny@ugecam-alsace.fr

The author from whom reprints can be obtained: A. Krasny-Pacini

If reprints are not available, this is stated on the title page.

Clinical trial registration number, if applicable: not applicable
Proposed criteria for appraising

Goal Attainment Scales used as
outcome measures in rehabilitation

Abstract

Goal Attainment Scaling (GAS) is a method for writing personalized evaluation scales in order to quantify progress toward defined rehabilitation goals. In published literature, GAS methodology is used with different levels of rigor, ranging from precisely written GAS scales, that ensure minimal bias, explicitly describing five levels of goal attainment; to subjective rating of goal attainment by adjectives such as “worse/better than expected”, which are transformed into a T-score, wrongly giving the reader the impression of a truly standardized, interval scale. A drawback of GAS methodology is that it is highly dependent on the ability of the GAS setting team/person to generate valid, reliable and meaningful scales, therefore reliability and validity of GAS scales are idiosyncratic to each study. The aims of this article were to: (1) increase awareness of potential sources of bias in GAS processes; (2) propose Goal Attainment Scaling quality appraisal criteria, allowing judgment of the quality of GAS methodology in individual rehabilitation studies; and (3) propose directions to improve GAS
implementation in order to increase its reliability and validity as a research measurement tool. Our proposed quality appraisal criteria are based on critical appraisal of GAS literature, and published GAS validity studies that have demonstrated that precision, validity and reliability can be obtained when using GAS as an outcome measure in clinical trials. We recommend that authors using GAS report accurately how GAS methodology was used based on these criteria.

Keywords
goal attainment scaling; outcome measures; goal setting; methodology; scale validity; scale reliability; standards; quality appraisal; guidelines

List of abbreviations
GAS: Goal Attainment Scaling
ICF: International Classification of Functioning
IRR: Inter-rater reliability
PMR: Physical Medicine and Rehabilitation
RoM: Range of motion
RCT: Randomized Controlled Trial
SMART criteria: Specific, Measurable, Attainable, Relevant, Time-specific
Introduction

Goal Attainment Scaling (GAS) is a method for writing personalized evaluation scales in order to quantify progress towards defined goals (see practical guidelines and literature reviews on GAS). Goal Attainment Scaling produces an individualized, criterion-referenced measure of a client’s goal achievement. Scores can be aggregated to quantify the extent to which a group of clients who are receiving the same type of intervention achieve their personalized rehabilitation goals. One GAS scale is written for each identified rehabilitation goal, with an emphasis on the client’s participation in goal selection when possible. Success of the intervention is then quantified on an ordinal scale, typically ranging from -2 (or -3) to +2.

GAS has therefore two intertwined components: (1) GAS methodology is a person-centered approach in rehabilitation that emphasizes collaborative goal setting with the establishment of goals and levels of progress that are meaningful to the client; (2) GAS is an outcome measure that can be used both in clinical work and research to assess the effectiveness of an intervention based on personally relevant goals. This paper focuses on the use of GAS as an outcome measure specifically for rehabilitation efficacy research. The reader is referred elsewhere to reviews of the literature on the clinical aspects of collaborative goal setting.

Writing personalized scales through GAS methodology is useful in measuring rehabilitation outcomes, and use of GAS methodology is expanding in research settings, especially in areas where standard scales do not adequately capture a study participant’s progress or when a standardized assessment does not exist to measure the construct. GAS methodology offers benefit in the provision of individualized, dependent variables, a critical characteristic for measuring rehabilitation effects. GAS allows use of the same 5-point scale
method for all clients and therefore aggregation of results independent of goal type. Further, the goal of rehabilitation is to improve clients’ activity and participation in natural contexts, but very few measures are designed to ecologically assess performance. By contrast, GAS allows the transformation of goals related to the International Classification of functioning (ICF) activity domains into participation goals in defined contexts where the activities occur. Feasibility of GAS has been shown across a variety of rehabilitation fields. GAS scales are sensitive to change when testing an intervention in rehabilitation. GAS characteristics in terms of safety, utility and responsiveness are therefore encouraging. However, in published literature, GAS methodology is used with different levels of rigor, ranging from precisely written GAS scales that ensure minimal bias, explicitly describing five levels of goal attainment; to subjective rating of goal attainment by adjectives such as “worse/better than expected”, which are transformed into a T-score, wrongly giving the reader the impression of a truly standardized, interval scale. Although the less rigorous form of GAS methodology can be convenient, useful, fast and practical to use in clinical practice, there is growing concern for its use as an outcome measure in clinical trials and mixed findings as to the reliability and validity of GAS as an outcome measure.

The aims of this paper are to (1) increase awareness of potential sources of bias in GAS processes; (2) propose Goal Attainment Scaling quality appraisal criteria, allowing judgment of the quality of GAS methodology in individual rehabilitation studies; and (3) propose directions to improve GAS implementation in order to increase its reliability and validity as a research measurement tool. This paper is not addressing use of GAS in clinical setting outwith research.

Methods
A literature search using PubMed database was conducted to ensure that our critical appraisal of the research was inclusive. The keywords “goal attainment scaling” and “rehabilitation OR therapy” were utilized to identify articles published between 1990 and 2014. The search returned 179 articles. Twelve articles were excluded because an abstract was not available or because the article was not written in English. A title and abstract review was conducted to identify those articles that evaluated GAS methodology as an outcome measure.

Included papers were: (1) Literature reviews on GAS; (2) GAS clinical guidelines; (3) Papers relating to GAS validity and reliability; (4) Papers relating to training in GAS. We purposefully included papers referring to fields outside Physical Medicine and Rehabilitation (PMR), that face the same challenges in evaluating treatment efficacy as rehabilitation does (especially cognitive interventions from the field of psychiatry and developmental disorders). Papers were excluded if they assessed only GAS feasibility or sensitivity to change/responsiveness, without references to its validity and reliability as an outcome measure. This yielded 36 relevant full text papers that were reviewed in order to identify bias in GAS and generate the quality appraisal criteria.

Potential sources of bias in GAS processes and published recommendations for constructing goal attainment scales.

Usual criticisms of how GAS methodology has been used include: (1) unknown clinimetric qualities of GAS scales used in a given study due to their idiosyncratic nature; (2) subjective scoring, especially if not all levels of the scale are formulated or if descriptions are not precise enough; (3) risk of choosing goals that are not clinically relevant or too easy/too difficult to attain and therefore do not represent a meaningful or realistic change in function; (4) ordinal (rather than interval) nature of GAS scales and the lack of equidistance
between GAS levels which cannot be controlled for; (5) the use of a T-score that uses subjective values, especially a subjective weighting of GAS scores and a $\rho$ coefficient assumed to be 0.3 which has not been confirmed in the literature.

A major drawback of GAS methodology is that it is highly dependent on the ability of the GAS setting team/person to generate valid, reliable and meaningful scales. It has even been proposed that GAS is more a measure of how adequately a therapist can foresee outcome than an outcome measure itself. A group of clients may show progress on their GAS scale due to a measurement error, on a GAS scale that is not reliable because of poor inter-rater reliability, too easy goals, unequal distances between GAS levels or use of subjective criteria for goal attainment. This issue has been raised by Ruble et al. 27: “If GAS scores are higher in the experimental conditions […] one could argue that the targeted outcomes as scaled using GAS were less difficult and easier for [clients] in the experimental group to achieve compared to the control group; that skills were written in more measurable terms and thus easier to be observed and coded in the experimental groups; or that the intervals between each scaled description were unequal and favored the experimental group.” (p3).

Because these potential biases can threaten reliability of results obtained through GAS, Kiresuk et al. 1,33 recommended the review of GAS scales by an independent third party, and even suggested that clients should be evaluated on two different sets of GAS scales, developed by two independent research groups (i.e. treatment success should be independent of how the goals were formulated) 1,34 to minimize bias. Although few publications address this demanding recommendation 35,36 it seems crucial that authors using GAS as a research outcome measure provide the reader with information on how the scales were generated and verified (and/or compared between groups on items that may impact
on GAS scoring as suggested by Ruble et al.\textsuperscript{27}), in order to provide information on reliability and validity. Some authors have found encouraging values of GAS reliability and validity\textsuperscript{25,15,37,22,35} (see recent systematic review by Vu et al\textsuperscript{8}). However the validity and reliability of GAS set by one team (especially an experienced one) does not presume that other GAS scales set by other teams, in other rehabilitation contexts, are valid and reliable\textsuperscript{27}. GAS clinimetric qualities depend mainly on how experienced the team is in GAS writing. Grant et al. reported the problems encountered when GAS are used by inexperienced teams, without an independent experienced judge checking the scales\textsuperscript{12}.

A series of criteria for writing GAS has been proposed\textsuperscript{38,1,2,35}: (1) each GAS level must be described accurately enough to allow a person who was not involved in the GAS-writing process to easily classify the client at one of the GAS levels described\textsuperscript{1}, with no “blank levels”\textsuperscript{39} (levels not precisely described, which content is inferred from adjacent levels); (2) each scale must represent a single dimension of change\textsuperscript{12}; (3) the levels must be measurable and thus defined in terms of observable behaviors\textsuperscript{6,9,40}; (4) the scales must correspond to goals that are important/meaningful to the client; (5) all the levels must be realistic and attainable (in particular, the +2 level must not correspond to an unexpected or miraculous goal attainment level)\textsuperscript{1}; (6) the time scale within which goals must be attained and scales must be scored should be defined in advance; (7) the inter-level differences in difficulty must all be the same\textsuperscript{41,42}, i.e. it must be as difficult to progress from −2 to −1, as from −1 to 0 or from 0 to +1, etc... and there should be no overlapping and no gap between the levels\textsuperscript{39}. Part of these criteria are reflected in the “SMART” acronym\textsuperscript{38} i.e. a goal should be Specific, Measurable, Achievable, Relevant and Time-determined. Although all authors acknowledge the need for GAS to be “SMART”, few report precautions taken to ensure GAS scales are...
Actually “SMART”, and virtually none assesses GAS quality when using it as an outcome measure.

Some authors proposed additional recommendations for GAS when it is used in research:
(1) including a training program \(43,25,15,44\); (2) establishing all goals prior to randomization\(45,46\) or blinding the goal-setter to the patient’s treatment/control status\(47\); (3) testing of interrater reliability for initial and post intervention GAS rating\(43,46\); (4) GAS scoring by a blind examiner\(46\) who is independent from the team that set the goals \(48,49,46,42\) and independent from the therapist providing intervention\(43,44\); (5) the use of “control goals” that are not targeted by the intervention\(6\); (6) evaluation of the patients on two different GAS scales developed by independent therapists (i.e. treatment success must be independent of how the goals were formulated) \(1,35,36\); (7) goal-setting by a group (rather than a single therapist or the patient alone), in order to avoid overly simple or unrealistic goals \(34\). To our knowledge, the impact of those recommendations on GAS validity and reliability has not yet been studied and few studies follow these guidelines. Their utility and applicability will be discussed in the discussion section of this paper.

**Proposed criteria for appraising Goal Attainment Scales**

Because GAS is a relevant and responsive outcome measure in rehabilitation research, but used with great variability that weakens the confidence in the results of trials that use this methodology, there is an urgent need for standards relating to GAS use in rehabilitation efficacy research. Our aim was therefore to propose GAS quality appraisal criteria, that would allow judgment of the quality of GAS methodology in individual rehabilitation studies, that could be used as guidelines to reduce bias, and strengthen GAS validity and reliability.
Based on our review of the literature, items for the quality appraisal were included if they met one of the following: (1) historically or traditionally recognized quality criteria (such as the “SMART” criteria and Kiresuk et al.'s\textsuperscript{1,33} rules for writing GAS scales); (2) criteria used by teams who obtained and published a good level of inter-rater reliability of their GAS data; (3) criteria used in rehabilitation trials to compare GAS quality across experimental groups; (4) items judged consensually by all authors of this paper as potential key candidates for increasing GAS validity and reliability (even in the absence of literature showing their impact on GAS clinimetric quality). Disagreements between authors on included items are developed in the discussion section.

In selecting criteria, the publications of two teams were particularly useful. In Steenbeek et al.'s methodology\textsuperscript{35,48,23}, eight GAS characteristics can be identified to ensure the construction of reliable scales: (1) all five levels of the GAS are precisely described; (2) GAS scales use objective and observable measures based on performance; (3) context of measurement is precisely described and factors that might influence performance are controlled for; (4) initial level is systematically verified after scale is set; (5) an independent, blind assessor scores GAS after intervention; (6) GAS data analysis respects the ordinal nature of GAS, using only raw scores and non-parametric statistics; (7) inter-rater reliability of GAS used in each study is reported; (8) teams are specifically trained to write reliable GAS scales (refer to Steenbeek et al.\textsuperscript{41} for an example of training).

Ruble et al. describe criteria for using GAS as a main outcome measure in a randomized controlled trial (RCT)\textsuperscript{27}. They suggest the following three key questions to ensure comparability of GAS in different experimental groups: “(a) are the goal and the associated benchmarks relative to each goal described in measurable terms that are comparable between groups (measurability criteria); (b) is the distance between each of the benchmarks
for each scale of equal intervals and comparable between groups (equidistance criteria); and (c) is the level of difficulty between the baseline or starting levels of performance and the targeted outcome goal comparable between groups (difficulty criteria)?” 27. In a recent controlled trial, all GAS scales were compared across the two experimental groups on these criteria using a three-point Likert scale, showing the feasibility of comparing GAS scales across groups in RCTs.

We propose 17 GAS quality criteria and these are presented in Table 1. They are broadly grouped into criteria that relate to the content validity of scales (4 items), the reliability of scale construction (4 items), reliability of scale rating (5 items), and an additional four items relating to training, examiner bias, statistical analysis and provision of a sample scale. If GAS is used in a controlled trial, we propose that GAS scales should not only be checked, but also compared between groups, similarly to the methodology proposed by Ruble et al. 27 for relevant items. For large trials, we propose that at least 20% of GAS scales be checked/compared.

GAS validity

GAS should be used when standardized assessment does not exist to measure the construct. Content validity of GAS scales is commonly thought to be high if the goal has been collaboratively set with the client wherever possible and this is the first criterion. Goals should be relevant and reflect clinically meaningful change, and this needs to be independently verified. To document the functional relevance of goals, the ICF domain that the goals reflect should be documented. Specificity is core to SMART goal-setting but with several definitions. Our specificity criteria relates to whether goals set specifically relate to
the intervention being tested - in research it is important to be able to articulate how a particular intervention will lead to achieving the desired goal.

GAS reliability

There are two types of reliability that are particularly important in GAS:

1. Reliability in the way the scale was constructed (i.e. even with an excellent inter-rater reliability, the scale may not be reliable because of non equidistant levels, erroneous starting pre-intervention levels, too easy goals/GAS levels, and an unspecified time frame for goals’ attainment influencing the relative difficulty of attaining a specific goal at the generic post-intervention assessment time point). The first four reliability criteria reflect these issues.

2. Reliability in scoring a given GAS scale. Measuring inter-rater reliability (IRR) provides a check on measurement accuracy. The following four items are thought to impact IRR: GAS scales where each level is not precisely described, that use subjective criteria for goal attainment with poor measurability or multidimensional scales, and which do not control for context of measurement, are likely to show lower IRR. Future research is needed to evaluate if respecting those criteria allows better IRR. An adequately measured and reported IRR could release authors from checking all GAS for those items (as training in GAS on those items is likely to generate scales with greater IRR). Because measuring their IRR is time consuming and requires an additional staff member, for large trials, we propose that at least 20% of GAS scales be tested for IRR (similar to the 20% of measurement criteria of N-of-1 trial standards and Ruble’s study).

Additional items

The final four criteria are included to further reduce potential for bias and increase confidence in GAS in research reports. They relate to: training of staff writing GAS scales;
independence of the person(s) evaluating goal achievement from those who set the goals; 
use of appropriate statistical analysis methods; and provision of examples of GAS scales in 
 research reports.

Discussion

The proposed criteria are indisputably challenging to meet. Most research using GAS 
does not meet all criteria. However it seems important to set out the highest possible 
criteria for all forms of research, even though few studies can meet all criteria. The proposals 
here are intended for research, where outcome measures must be valid and reliable to 
prevent erroneous conclusions about the effectiveness of an intervention. They do not 
discredit less rigorous but more user-friendly, more practical and less time-consuming uses 
of GAS in clinical practice.

Controversial criteria

Should GAS use be restricted to collaboratively set goals? In the literature, goals are 
often set in collaboration between client/family and therapist, but may also be 
chosen by the therapist alone, or by the client/family alone. Initially, the GAS 
methodology was invented to assess any goal-directed enterprise, including the functioning 
of a Crisis Intervention Center or a hospital-based pharmacy project. Initially goals could 
therefore be chosen by any professional, with or without participation of the client. In the 
last twenty years, as rehabilitation moved towards a more person-centered approach, GAS 
has become increasingly used as a method for collaborative goal setting, as well as an 
outcome measure. Literature on brain injury developed guides and methods of
collaborative goal setting, and linked use of GAS as an outcome measure with active client participation in goal setting. However GAS may be valuable in PMR domains where client participation in goal selection is not possible (e.g. patients with minimally conscious state, clients with severe intellectual and behavioral impairment...) or is not essential (e.g. early interventions after stroke to prevent contractures and shoulder pain; motor development in infants with attainment of developmental stages that may not be relevant for the family immediately but that are believed to be crucial for future development and future more functional goals...). GAS may also be used to assess effectiveness of an intervention at a health provider level (e.g. goal of reducing pressure ulcer incidence/need for surgery after a group therapeutic education in a spinal cord injury unit). Because collaborative goal-setting is time-consuming for therapists and cognitively demanding for clients (especially those with brain injury), therapists may use collaborative goal setting for some goals, while choosing themselves goals of other domains that are indisputably useful for this client and the focus of the intervention being tested. Therefore we propose here a less restrictive requirement (in the collaborative goal setting item), allowing therapist-chosen goals. In all cases, authors should report if goals were not collaboratively set and provide reasons of their choice.

Most often, GAS scales are set for goal-focused rehabilitation where the goal is directly trained and GAS represents the degree of progress towards a goal. However it has also been proposed that GAS could be a measure of generalization: i.e. after training relating to a cognitive function (e.g. executive functions, memory), the use of GAS scales can help assess if training lead to gains in daily life (e.g. GAS relating to activities relying on executive functions such as preparing a schoolbag, GAS relating to memory such as taking medicines on time), without specifically addressing these goals. A valuable approach is to use both a set
of trained goals (and corresponding GAS) and a set of untrained goals (and corresponding GAS) and then to focus the intervention on training the former while using the latter as an untrained ecological generalization measure. Further it is has been proposed that “control” goals (and related GAS scales) that are not expected to show progression are used, in order to demonstrate the specificity of an intervention (i.e. the client does not just progress on all goals because of general cognitive stimulation or goal-driven motivation, but progresses on the specific goal that is trained or that relates to the trained function needed to achieve an untrained generalization goal).

Authors should report the types of goals chosen in their study, using the ICF. GAS types vary considerably between studies, and do not always measure a meaningful goal, but may remain in the body structure and function (gait pattern, range of motion). A clear demonstration of functional benefit to the client in terms of activity and participation is increasingly required in order to show an intervention is effective, and GAS should relate to activity and participation domains as much as possible. When using GAS for body structures, there is a risk that GAS be used as a way of getting round the need for standardized measures. A methodological error often seen with GAS is to convert existing (or even standardized) scales into GAS scales. This is done for two practical reasons: (1) to help measure a particularly complex goal (e.g. if cognitive restructuring [...] is an important treatment component, it may be advantageous to include a standardized pain beliefs scale as a goal area; (2) to transform a meaningless number into a relevant and meaningful goal; (3) in order to obtain the same outcome measure for all clients. For example for botulinum toxin treatment, the goal of one client may be to decrease pain, measured on a visual analogue scale, then transformed into a GAS scale depending on the pain level considered as meeting the goal of a treatment; the goal for another client may be to
decrease equinus (measured in degrees) in a gait analysis laboratory and the range of motion is transformed into a GAS scale presenting a range of ankle positions as the goal. However such a conversion is done at the expense of losing the linearity of the original measure and although very useful in clinical practice, it is scientifically acceptable only if data is analyzed as truly ordinal (therefore not using T-scores, nor means nor any arithmetical operation). Can GAS levels be equidistant and GAS data be interval in form? This is possible if a calibration by Rasch Analysis on an “item bank” is carried out as proposed by Tennant et al., but at the expense of losing GAS adaptability to any goal. Otherwise, GAS levels are very unlikely to be interval despite all precautions, corrections and comparisons used to minimize level inequality bias. Therefore ordinal interpretation of GAS, using rank tests (see Steenbeek et al for an example) and excluding all arithmetic procedures on GAS scores seems the most reasonable option for GAS data analysis. It is indisputably difficult to set 5 equidistance GAS levels but simple rules can be postulated to facilitate choice of levels (both for goal setters and the external judge who checks GAS levels and compares equidistance between two groups): (1) avoid setting “half levels” (e.g. -0.5 as proposed by Turner Stokes); (2) have all clients start from the same initial score and therefore have all clients assessed on the same number of levels of goal attainment (see for a discussion on advantages of scoring initial level at -2 or at -1).

What type of staff training should be required in order to use GAS in a research protocol? Kiresuk, Smith and Cardillo had proposed that a minimum of one-year experience is required to develop relevant and realistic scales. Basic knowledge of GAS, and experience in collaborative goal setting is not sufficient, as shown by Grant et al. Reading a practical guide may be sufficient for GAS clinical use but not for studies aiming at producing valid
and reliable GAS. A number of authors\textsuperscript{70,41} propose practical training that is largely based on formulating and correcting GAS scales, based on clients’ real goals. We recommend that such a practical training be used. This should also raise awareness on GAS quality criteria the study will be judged on, and make GAS scoring easier after the intervention. The quality appraisal here proposed could be the focus of such training.

Challenges in GAS methodology

GAS was intended to be a person-centered measure. Is there a risk of losing its person-centered nature when trying to meet measurement criteria for controlled efficacy trials? The risk of insisting upon an observable and measurable goal is to exclude family and client appreciation of goal attainment. Appreciation of goal attainment by clients may be subjective and is affected by a series of factors (self-awareness, denial, memory, high involvement in goal pursuit) that can bias the perceived attainment of a goal; however goal identification and measuring its attainment should be person-centered. The challenge for clinicians and researchers is therefore to understand and analyze the client’s (and/or his/her families) goal and transform it into an observable, objective, performance-based measure that can then be used to discuss goal attainment with the client. For example a general goal of “improving my memory” expressed by a client can be transformed into a GAS scale that measures memory functioning in real life situations agreed with the client (e.g. number of medicines taken on time without prompting; number of failed-to-deliver messages in one week…). Scoring GAS based on simple interview should be avoided when possible and creativity used to link subjective goals in difficult domains to more objective goal attainment indicators. For example for a GAS scale on anger management or use of social skills, rather than asking the client to recall how often she/he felt he could cope with his anger or
effectively use social/language skills, an effort should be made to choose indirect indicators of goal attainment such as a review of controlled versus overt anger at the end of a day by the client or proxy rating of social/language skill (how many times he/she initiated/contributed to conversation or was understood) after regular naturally occurring events (outing with friends, family dinner). It is probably the most challenging part of GAS methodology but it has been demonstrated that it is feasible to have GAS that are both person-centered in the choice of goal and objective in the formulation of GAS scales (see Steenbeek et al. for an example). N-of-1 trial literature faces the same challenges and offers growing ingenuous methods for assessment of domains not directly accessible to classical performance-based objective measure. These methods include use of smartphone reminders to self-assess goal attainment at regular periods to decrease memory bias (see for an example), use of naturally occurring situations monitored in real life by family/proxy to decrease self-awareness bias (see for an example), use of objective behavioral measures that are thought to reflect the underlying psychological (e.g.: happiness) or cognitive (e.g.: functioning at school) target goals. Future research should extend “goal menus” such as those proposed by Turner-Stokes et al. to more challenging domains (such as goals relating to social and psychological functions, as well as goals focusing on performance in ecological setting), using also N-of 1 trial literature. When goal attainment indicators are ecological and monitored by the client or proxies, calculating an IRR is impossible (or difficult) and GAS should be checked for items: precise description of all GAS levels, measurability, unidimensionality and context of measurement to increase reliability in scoring. A reasonable compromise between scientific rigor and person-centered approach could be to have for each client at least one ecological client-centered GAS (with the risk of being less
reliable) and at least one performance-based GAS (with the risk of being less ecological). In all cases, Kiresuk recommended to “anchor scale points with behavioral or other evidence that will be meaningful to the client and readily scored by the rater”\(^{33}\) (p 31).

Limitations

The present article has a number of limitations. Choice of included criteria was not based on a consensual agreement of all major teams using GAS in research but of four teams, from three different countries. The aim of this paper was to raise awareness about the variability of GAS use in published research and the need to build, in future, a consensus on the use of GAS in efficacy research in rehabilitation. Although it may be viewed as a limitation, we purposefully did not validate externally an appraisal score, so that the present guide acts only as a starter for discussion and not a validated tool imposed on other teams that were not included in the writing of the paper.

Conclusion

Goal Attainment Scaling has the potential to be sensitive to change following treatment and applicable across divergent domains of rehabilitation, making it a useful rehabilitation efficacy research outcome measure. However, GAS is used in studies with variable rigor that impacts its validity and reliability, and therefore reduces the confidence one can have in the results of a trial using GAS as an outcome measure. Clinimetric qualities of GAS are highly dependent on the way GAS scales are written and therefore clinicians, researchers and reviewers cannot rely on published studies of metrological qualities of GAS obtained in different research studies. Researchers should be aware of the risk of bias related to the use of imprecisely written GAS scales in research and make all possible efforts to minimize this
bias by constructing high quality GAS scales for their clients, following recommendations described here and previously published examples\textsuperscript{27,48}. There is a need to develop in the future a GAS quality appraisal score, similarly to standards used in other rehabilitation controversial fields such as N-of-1 trials\textsuperscript{51}. In this paper, we propose 17 criteria for appraising GAS quality in trials using GAS as an outcome measure. We recommend that authors using GAS report accurately how GAS methodology was used based on these criteria. The reliability of these criteria needs to be established, but in the meantime we invite comment and discussion of these proposals as we move towards a consensus on standards for use of GAS in rehabilitation research.

References


17. Hazard RG, Spratt KF, McDonough CM, Carayannopoulos AG, Olson CM, Reeves V, et al. The impact of personal functional goal achievement on patient satisfaction with progress one year following completion of a functional restoration program for chronic


673
674  69.  Turner-Stokes L, Williams H. Goal attainment scaling: a direct comparison of alternative
676
678  Goal attainment scaling: Applications, theory, and measurement. Hillsdale, NJ,
680
681  71.  Arco L. Neurobehavioural treatment for obsessive-compulsive disorder in an adult with
683
684  72.  Baldwin VN, Powell T. Google Calendar: A single case experimental design study of a
686
690
691  74.  Feeney TJ, Ylvisaker M. Context-sensitive behavioral supports for young children with
693
694
695  75.  Feeney T, Ylvisaker M. Context-sensitive cognitive-behavioural supports for young
697
698  76.  Ashford S, Jackson D, Turner-Stokes L. Goal setting, using goal attainment scaling, as a
699  method to identify patient selected items for measuring arm function. Physiotherapy.
700  2015;101:88–94.
<table>
<thead>
<tr>
<th>Quality appraisal items</th>
<th>Item description</th>
<th>Examples of reported criteria, extracted from rehabilitation effectiveness studies and/or published methods that satisfy criteria.</th>
<th>Potential for bias arising from failure to report criteria and/or utility of reporting the criterion.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content Validity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Collaborative goal setting | The client/family is included in goal selection, when possible and appropriate. If goals are therapist-chosen, they rely on a comprehensive assessment (and when possible, a client/family interview), identifying key target domains for intervention. | “GAS was used to assess functional and participation changes from both a parent and therapist perspective” 
“Using a semi-structured interview [...] 3 performance goals were identified at baseline by parents and child with the research physical therapist. The same 3 goals were then structured for GAS through semi-structured interview and by collaboration between research therapist, parent and child.” | Collaborative goal setting allows evaluation of intervention efficacy for personally meaningful goals, rather than generic goals and is a core component of GAS methodology. Involvement of the client in goal setting is considered to increase the likelihood that the intervention has direct impact on client’s daily life. |
| **Relevance/ importance** | GAS scales have been verified by an external judge to check for the relevance of chosen goals and to check if GAS levels represent clinically meaningful change. | In a study of infants with motor delays, Palisano used a 5 point scale to assess: (1) importance of goals for motor development ranging from “unimportant or inappropriate” to “important for development and for function”; (2) extent each level represents an important progress based on number of paired levels that represent important change (“none” to “all four paired levels”). Cardillo proposed a 5 point scale ranging from “1: No relevance” to “5: Total relevance”. | If the target goal is unimportant to the client, irrelevant for function or does not correspond to a clinically meaningful change, progress on the GAS scale has no clinical relevance. At the extreme, an intervention could be proven to be effective, by writing clinically and personally irrelevant goals but showing statistically significant progress on the corresponding GAS scale. |
| **ICF classification of goal types** | GAS themes correspond to functional domains. Authors report the ICF domain the GAS relate to. | In a study of botulinum toxin effectiveness, Turner-Stokes et al. report exhaustively the types of goals for treatment, categorizing them into ICF domains (Body function: 46 GAS, including 12 GAS on passive movement/range; Activities and Participation: 119 GAS, divided into: Upper limb activities: 30 GAS; Mobility: 11 GAS; Self care: 57 GAS; Domestic and Community: 21 GAS). Phillips et al. report precise examples of target goals in different ICF domain (e.g.: “driving for 40 minutes without feet going floppy”; “standing in supermarket queue for 3 minutes without support”). | If GAS scales assess change in body structures (e.g. RoM, spasticity), the reader may wrongly conclude that the intervention had an impact on meaningful activity and participation, because most readers associate GAS with functional daily life goals. It is therefore crucial that authors report the proportion of goals in each ICF domain, especially if some of goals do not correspond to functional domains. |
| **Specificity** | GAS scales have been verified by an external judge to check for specificity to the aim of the intervention. If GAS is used as a generalization measure to untrained goals, GAS should be specific to the function the goal is relevant to. | In an intervention for executive dysfunction in children, a specific goal of intervention was to manage a cooking recipe unaided, which was trained on different recipes; a generalization goal was to be able to prepare school bag which was not trained (but stepwise processing taught was expected to generalize to this untrained goal). Although important to the children, goals such as “have more friends”, were not included as they were not specific to the aim of goals that are relevant to the client, but unrelated to the specific intervention, are unlikely to show progress and may erroneously lead to the conclusion that the intervention is not effective (this is especially a risk in replication studies). | Null |
intervention is aiming to improve. Intervention, but they might have been used as “control” goals, not expected to be attained.

<table>
<thead>
<tr>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reliability of scale construction</strong></td>
</tr>
</tbody>
</table>

**Equidistance of levels**

GAS scales have been verified by an external judge to check if difficulty from one level to the next is roughly equal. 

Equidistance of GAS levels was assessed and compared statistically between two experimental groups using a 3 point scale by Ruble et al. 8,9: “1: None or only one of the descriptions are equilibrated appropriately in reference to the goal; 2: Two of the descriptions are equilibrated appropriately in reference to the goal; 3: All of the descriptions relative to the goal are equilibrated and scaled appropriately”.

If GAS scores are higher in the experimental conditions, one could argue that the intervals between each scaled description are unequal and favor the experimental group. This problem is particularly serious if parametric statistics and T-scores are used.

**Pre-intervention performance**

Pre-intervention performance has been verified and corresponds to initial level described in the scale. Pre-intervention score is comparable across groups (same number of clients starting from -2 and -1).

“At first baseline, GAS were created based on parent’s, teacher’s and school assistant’s concerns. After two months (second baseline), the paper versions of the scales were scored by parents, teachers and school assistants who were not aware that the intervention had not started yet. Their answers allowed readjustment of the scales, through the following rules: Scales that scored 0 were reformulated in order to have the pre-intervention level (measured at second baseline) corresponding to -1 by fixing more challenging 0, +1 and +2 scores. Scales scoring -2 or -1 were not reformulated. Scales scoring +1 or +2 pre-intervention were removed as the goal seemed attained without intervention or unreliably scored due to potential enthusiasm effect and motivation driving perceived change independently from intervention that had not started”.

GAS scales are constructed uniquely for each client, according to his/her initial level in relation to the target goal. If pre-intervention level is not verified, the whole scale may be unreliably constructed (erroneous starting point of the scale generating inappropriate next levels).

**Attainability/difficulty**

GAS scales have been verified by an external judge to check for their difficulty/attainability.

Ruble et al. compared GAS scales for difficulty in a trial of cognitive intervention: “1: Skill is very close to what the child is already described as able to perform (very easy); 2: The child is able to perform the skill in limited ways compared to what is written in the objective (limited people, prompts, or places [...]); 3: The child is unable to perform skill with anyone, anywhere, or with any prompts compared to what is written in the objective (difficult)”.

In a study of infants with motor delays, Palisano asked experts to decide which of the 5 GAS level the child was most likely to achieve after 3 months, aiming at a maximum of 0 and a minimum of extreme -2 and +2 scores, if GAS levels were decided correctly. Cardillo reported that realism of the expected level of outcome
“for each goal, for this patient, at this time, in this mental health service” was assessed by a scale ranging from 1: much too difficult to 5: much too easy. He also reviews other methods that use 3 or 5 point scales using “pessimistic/realistic/optimistic” terms. Such scales could be easily used in PMR to compare attainability/difficulty of goals between two groups.

<table>
<thead>
<tr>
<th>Time-specificity</th>
<th>Authors specify if/how longer-term goals were adapted to the specific time frame of the research study. In the case of multiple assessment, authors specify which assessment was taken as the target moment for goal achievement.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In the RCT of Lowe et al., children were evaluated at 13 different time points. It is not reported which assessment point was taken as the reference to choose the 0 level (level that will most probably be attained after intervention) – criterion unmet.</td>
</tr>
<tr>
<td></td>
<td>Goal difficulty across experimental groups may have been unequal, at a given assessment point, due to differences in time-frames for goal completion.</td>
</tr>
</tbody>
</table>

**Reliability of scale rating**

<table>
<thead>
<tr>
<th>Inter-rater reliability (IRR)</th>
<th>Inter-rater reliability of GAS scales is reported.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In a study of Steenbeek et al. in cerebral palsy, inter-rater reliability was reported based on two judges using video-taped performance of each goal.</td>
</tr>
<tr>
<td></td>
<td>Reliability of GAS set by one team (especially an experienced one) does not presume that other GAS scales set by other teams, in other rehabilitation contexts, are reliable. Therefore IRR should be reported for the specific GAS scales generated in each study.</td>
</tr>
</tbody>
</table>

**Criteria affecting IRR**

<table>
<thead>
<tr>
<th>Precise description of all levels</th>
<th>Five GAS levels have been precisely described pre-intervention for each scale.</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Goal: Reducing weekly shopping expenditure”</td>
<td>Total weekly food/household shopping expenditure = +2: less than $42.99, +1: $46.99–$43.00, 0: $54.00–$47.00, -1: $63.00–$54.01, -2: greater than $63.00.”</td>
</tr>
<tr>
<td></td>
<td>When all levels of the scales have not been precisely decided and described prior to intervention, authors often use adjectives such as “worse than expected”, “much better than expected”, to score goal attainment. This is a subjective appreciation that may be useful clinically but is too imprecise to objectively determine intervention efficacy.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurability</th>
<th>GAS scales have been verified by an external judge to check for measurability. Subjective and general goals are transformed into more objective and measurable goal attainment indicators.</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;good measurability&quot;</td>
<td>Observable and objective performance with specified task (e.g.: A child’s goal to fall less is assessed through “an obstacle course including jumping and quick changes of walking direction. The therapist encourages [the child] to complete the course within 3min. Instruction “Walk the obstacle course fast and don’t fall”; GAS levels: -1: falls 3 times, 0: falls 2 times […]”).</td>
</tr>
<tr>
<td>&quot;unclear measurability&quot;</td>
<td>Subjective criteria, or scored based on interviews rather than direct observation of performance (e.g.: “I am able to express opinions and feelings two times or more per week […]”, with no self-assessment method specified).</td>
</tr>
</tbody>
</table>

A goal that is not measurable will yield subjective scores, biased by clients’ or therapists’ feelings/ state of mind at the moment of scoring rather than a reliable measure of goal attainment. See Ruble et al. for an example of assessment of measurability of
<table>
<thead>
<tr>
<th><strong>Unidimensionality</strong></th>
<th>GAS scales have been verified by an external judge to check for unidimensionality.</th>
<th>Social and cognitive goals(^*).</th>
<th>Non-unidimensional goals are impossible to score as progress on one dimension may not be accompanied by progress on another dimension and generate situations where GAS cannot be scored (see Grant(^{15}) for an example). Bi (Multi) dimensional GAS’s should be split into two (or more) unidimensional GAS prior to intervention start.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context of measurement</strong></td>
<td>Context of performance measurement is clearly defined (prompts, cueing, support, amount of help/guidance, location...) and is controlled for during GAS rating. OR Changes in context are carefully manipulated across the GAS levels, with one change per level at a time(^{16}).</td>
<td>Example of a non-unidimensional scale: “-1: I use 0–1 coping skills consistently and feel depressed and angry more than 40% of the time.; 0: I use 2–3 coping skills consistently and feel depressed and angry 25–40% of the time [...]”(^{11}) Example of a truly unidimensional scale: “... -1: Manages to eat a bowl of mashed potatoes unaided but takes more than 15 minutes.; 0: Eats a bowl of mashed potatoes in 11 to 15 minutes...”(^{14})</td>
<td>Context of measurement influences performance on a given target goal (environment, fatigue, help provided...). These factors must be controlled in order to increase GAS scoring reproducibility.</td>
</tr>
<tr>
<td><strong>Training</strong></td>
<td>Researchers setting the GAS with the client and verifying GAS have received training in writing GAS, have practiced GAS writing, are aware of potential sources of bias in GAS, and are experienced in the goal domain/population.</td>
<td>Although training in GAS writing is reported as being important(^{18}), and successful training methods have been published(^{17}), most studies do not report on therapists’ training. Some studies report to which practical GAS guide(^{16,19}) they refer to, but without mention of training(^{20,21}). Those mentioning training do not explain the type of training (e.g.: “Experienced pediatric occupational therapists were trained in and completed the GAS collaboratively with the families, thus enhancing the reliability of the GAS”(^{22})).</td>
<td>Given the numerous and complex potential sources of bias in GAS processes, a team that is not experienced in using GAS methodology is unlikely to produce valid and reliable GAS. Further, a team without specific experience in the goal domain or the specific population with whom the intervention is tested, will have difficulty in predicting what can be attained in a given time-frame, even if experienced in GAS methodology in another domain (risk unrealistic goals, unequal difficulty across clients, irrelevant goals to the specific population...).</td>
</tr>
<tr>
<td><strong>Examiner bias</strong></td>
<td>The person scoring the GAS at the end of the intervention is independent from the team who set the GAS (and independent from the team that provided the intervention although the latter is not a GAS specific criterion).</td>
<td>e.g.: “Goals were chosen and set before the patient was allocated to a group [...] goal attainment was scored by an independent assessor at post-treatment and at follow-up.”(^{20}).</td>
<td>e.g.: “The therapist-GAS was scored from video by blind evaluators. The parent-GAS was scored by two blinded occupational therapists.”(^{1}).</td>
</tr>
</tbody>
</table>


respected when goals are client/family chosen when GAS is an outcome measure in research, (in contrast with clinical practice, where GAS scoring by the client may be relevant and appropriate).

**Statistical analysis**

Ordinal nature of GAS scales is preserved using non-parametric statistics (rank tests, medians, boxplots).

*Example:* "It was decided not to use the popular T score in order to preserve the ordinal nature of the data [...] group effects were demonstrated by testing the difference between all medians, [...] using a two-tailed Wilcoxon signed ranks test".  

The performance of arithmetic operations such as T-scores on ordinal data is scientifically not valid and should be discouraged, as it yields erroneous interpretation of data. In GAS, the problem is multiplied by characteristics of the T-score formulae (unknown true value of ρ, T-score variation according to the number of goals per client even at equal degree of attainment, highly subjective weighting of goals which, although clinically meaningful, introduces further potential arithmetic incoherence in the final T-score).

**Example of GAS**

One (or more) example of a typical GAS full scale, extracted from the trial, is provided.

A list of chosen goals is reported.

Some authors provide an example of full GAS scale in the paper in the methods or results section.

Examples of goal types can be given (1) by providing examples of goal in the paper; (2) by reporting all chosen goals in appendix; (3) by reporting goal type and frequency of each type without providing an exhaustive goal list.

Providing examples allows the reader a quick judgment/idea of goal type, precision of goal and levels description, measurability, unidimensionality of GAS. The lack of GAS examples contributes to make GAS seem like an abstract outcome measure: unlike standardized scales, the reader cannot build a representation of the target goals of the intervention. Therefore, reporting all goals in an appendix and providing example(s) of full GAS scales (representative of different domains measured) should be encouraged to increase interpretability.

**TABLE 1: Goal Attainment Scaling (GAS) methodology quality appraisal for rehabilitation efficacy studies**

Criteria are grouped in accordance to the clinimetrics they mostly impact, although some items may impact both reliability and validity.

If GAS is used in a controlled trial, GAS scales should not only be checked but also compared between groups, similarly to the methodology proposed by Ruble et al. For large trials, we propose (as for IRR) that at least 20% of GAS scales be checked.

**ICF: International Classification of Functioning; IRR: inter-rater reliability; RoM: Range of motion**


