HAL
open science

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:

Agata Krasny-Pacini, Jonathan Evans, Mckay Moore Sohlberg, Mathilde Chevignard. Proposal of criteria for appraising Goal Attainment Scales used as outcome measures in rehabilitation research. Archives of Physical Medicine and Rehabilitation, 2016, 97 (1), pp.157-170. 10.1016/j.apmr.2015.08.424 . hal-01203473

HAL Id: hal-01203473

https://hal.sorbonne-universite.fr/hal-01203473v1

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.

ACCEPTED MANUSCRIPT

Title page

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-Pacini[1,5,6], MD

Jonathan Evans[2], PhD

McKay Moore Sohlberg[3], PhD

Mathilde Chevignard[4,5], 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

ACCEPTED MANUSCRIPT

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)

ACCEPTED MANUSCRIPT

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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

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

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

ACCEPTED MANUSCRIPT

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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Introduction

Goal Attainment Scaling (GAS) [1] is a method for writing personalized evaluation scales in order to quantify progress towards defined goals (see practical guidelines[2,3,4,5] and literature reviews on GAS [6,7,8,9,10]). 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 [10,9,11,12].

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

ACCEPTED MANUSCRIPT

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 [13,14]. Feasibility of GAS has been shown across a variety of rehabilitation fields[15,16, 17,18,19]. GAS scales are sensitive to change when testing an intervention in rehabilitation[20,21,15,16,22,23]. 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[24,25] and mixed findings as to the reliability[24,26] and validity of GAS as an outcome measure[25].

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

ACCEPTED MANUSCRIPT

A literature search using PubMed data base 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[25]; (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[27] and therefore do not represent a meaningful or realistic change in function; (4) ordinal (rather than interval) nature of GAS scales[28] and the lack of equidistance

ACCEPTED MANUSCRIPT

between GAS levels which cannot be controlled for[24]; (5) the use of a T-score that uses subjective values, especially a subjective weighting of GAS scores and a ρ coefficient assumed to be 0.3 which has not been confirmed in the literature[9,28,6].

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 [6,29,30,31,32]. 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 measureable 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[1] (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

ACCEPTED MANUSCRIPT

130 on GAS scoring as suggested by Ruble et al.[27]), in order to provide information on reliability
131 and validity. Some authors have found encouraging values of GAS reliability and validity
132 [25,15,37,22,35] (see recent systematic review by Vu et al [8]). However the validity and reliability of
133 GAS set by one team (especially an experienced one) does not presume that other GAS
134 scales set by other teams, in other rehabilitation contexts, are valid and reliable [27]. GAS
135 clinimetric qualities depend mainly on how experienced the team is in GAS writing. Grant et
136 al. reported the problems encountered when GAS are used by inexperienced teams, without
137 an independent experienced judge checking the scales[12].
138
139 A series of criteria for writing GAS has been proposed [38,1,2,35] : (1) each GAS level must be
140 described accurately enough to allow a person who was not involved in the GAS-writing
141 process to easily classify the client at one of the GAS levels described[1], with no "blank
142 levels"[39] (levels not precisely described, which content is inferred from adjacent levels); (2)
143 each scale must represent a single dimension of change[12]; (3) the levels must be measurable
144 and thus defined in terms of observable behaviors[6,9,40]; (4) the scales must correspond to
145 goals that are important/meaningful to the client; (5) all the levels must be realistic and
146 attainable (in particular, the +2 level must not correspond to an unexpected or miraculous
147 goal attainment level)[1]; (6) the time scale within which goals must be attained and scales
148 must be scored should be defined in advance; (7) the inter-level differences in difficulty must
149 all be the same[41,42], i.e. it must be as difficult to progress from –2 to –1, as from –1 to 0 or
150 from 0 to +1, etc… and there should be no overlapping and no gap between the levels[39]. Part
151 of these criteria are reflected in the ''SMART'' acronym [38] i.e. a goal should be Specific,
152 Measurable, Achievable, Relevant and Time-determined. Although all authors acknowledge
153 the need for GAS to be "SMART", few report precautions taken to ensure GAS scales *are*

ACCEPTED MANUSCRIPT

154 *actually* "SMART", and virtually none assesses GAS quality when using it as an outcome
155 measure.

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

169 Proposed criteria for appraising Goal Attainment Scales

170 Because GAS is a relevant and responsive outcome measure in rehabilitation research,
171 but used with great variability that weakens the confidence in the results of trials that use
172 this methodology, there is an urgent need for standards relating to GAS use in rehabilitation
173 efficacy research. Our aim was therefore to propose GAS quality appraisal criteria, that
174 would allow judgment of the quality of GAS methodology in individual rehabilitation studies,
175 that could be used as guidelines to reduce bias, and strengthen GAS validity and reliability.
176

ACCEPTED MANUSCRIPT

177 Based on our review of the literature, items for the quality appraisal were included if
178 they met one of the following: (1) historically or traditionally recognized quality criteria (such
179 as the "SMART" criteria and Kiresuk et al.'s[1,33] rules for writing GAS scales); (2) criteria used
180 by teams who obtained and published a good level of inter-rater reliability of their GAS data;
181 (3) criteria used in rehabilitation trials to compare GAS quality across experimental groups;
182 (4) items judged consensually by all authors of this paper as potential key candidates for
183 increasing GAS validity and reliability (even in the absence of literature showing their impact
184 on GAS clinimetric quality). Disagreements between authors on included items are
185 developed in the discussion section.

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

196 Ruble et al. describe criteria for using GAS as a main outcome measure in a randomized
197 controlled trial (RCT)[27]. They suggest the following three key questions to ensure
198 comparability of GAS in different experimental groups: "(a) are the goal and the associated
199 benchmarks relative to each goal described in measurable terms that are comparable
200 between groups (measurability criteria); (b) is the distance between each of the benchmarks

ACCEPTED MANUSCRIPT

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[50], 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.

INSERT TABLE 1 ABOUT HERE

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

ACCEPTED MANUSCRIPT

224 the intervention being tested - in research it is important to be able to articulate how a
225 particular intervention will lead to achieving the desired goal.

226 GAS reliability

227 There are two types of reliability that are particularly important in GAS:
228 (1) Reliability in the way the scale was constructed (i.e. even with an excellent inter-rater
229 reliability, the scale may not be reliable because of non equidistant levels, erroneous starting
230 pre-intervention levels, too easy goals/GAS levels, and an unspecified time frame for goals'
231 attainment influencing the relative difficulty of attaining a specific goal at the generic post-
232 intervention assessment time point). The first four reliability criteria reflect these issues.
233 (2) Reliability in scoring a given GAS scale. Measuring inter-rater reliability (IRR) provides a
234 check on measurement accuracy. The following four items are thought to impact IRR: GAS
235 scales where each level is not precisely described, that use subjective criteria for goal
236 attainment with poor measurability or multidimensional scales, and which do not control for
237 context of measurement, are likely to show lower IRR. Future research is needed to evaluate
238 if respecting those criteria allows better IRR. An adequately measured and reported IRR
239 could release authors from checking all GAS for those items (as training in GAS on those
240 items is likely to generate scales with greater IRR). Because measuring their IRR is time
241 consuming and requires an additional staff member, for large trials, we propose that at least
242 20% of GAS scales be tested for IRR (similar to the 20% of measurement criteria of N-of-1
243 trial standards [51] and Ruble's study[27]).

244 Additional items

245 The final four criteria are included to further reduce potential for bias and increase
246 confidence in GAS in research reports. They relate to: training of staff writing GAS scales;

ACCEPTED MANUSCRIPT

247 independence of the person(s) evaluating goal achievement from those who set the goals;
248 use of appropriate statistical analysis methods; and provision of examples of GAS scales in
249 research reports.

250 Discussion

251

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

259 Controversial criteria

260 Should GAS use be restricted to collaboratively set goals ? In the literature, goals are
261 often set in collaboration between client/family and therapist[52 53 54 55 56] , but may also be
262 chosen by the therapist alone[57 58], or by the client/ family alone [57 59]. Initially, the GAS
263 methodology was invented to assess any goal-directed enterprise, including the functioning
264 of a Crisis Intervention Center[34] or a hospital-based pharmacy project[34]. Initially goals could
265 therefore be chosen by any professional, with or without participation of the client. In the
266 last twenty years, as rehabilitation moved towards a more person-centered approach[60], GAS
267 has become increasingly used as a method for collaborative goal setting, as well as an
268 outcome measure. Literature on brain injury[61] developed guides and methods of

ACCEPTED MANUSCRIPT

collaborative goal setting [32 62 9 10], 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[63]: 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

ACCEPTED MANUSCRIPT

293 of trained goals (and corresponding GAS) and a set of untrained goals (and corresponding
294 GAS) and then to focus the intervention on training the former while using the latter as an
295 untrained ecological generalization measure. Further it is has been proposed that "control"
296 goals[6] (and related GAS scales) that are *not* expected to show progression are used, in order
297 to demonstrate the specificity of an intervention (i.e. the client does not just progress on all
298 goals because of general cognitive stimulation or goal-driven motivation, but progresses on
299 the specific goal that is trained or that relates to the trained function needed to achieve an
300 untrained generalization goal).

301 Authors should report the types of goals chosen in their study, using the ICF. GAS types
302 vary considerably between studies, and do not always measure a meaningful goal, but may
303 remain in the body structure and function (gait pattern[64], range of motion[29]). A clear
304 demonstration of functional benefit to the client in terms of activity and participation is
305 increasingly required in order to show an intervention is effective, and GAS should relate to
306 activity and participation domains as much as possible. When using GAS for body structures,
307 there is a risk that GAS be used as a way of getting round the need for standardized
308 measures. A methodological error often seen with GAS is to convert existing (or even
309 standardized) scales into GAS scales. This is done for two practical reasons: (1) to help
310 measure a particularly complex goal (e.g. if cognitive restructuring […] is an important
311 treatment component, it may be advantageous to include a standardized pain beliefs scale
312 as a goal area[47], p62); (2) to transform a meaningless number into a relevant and meaningful
313 goal; (3) in order to obtain the same outcome measure for all clients. For example for
314 botulinum toxin treatment, the goal of one client may be to decrease pain, measured on a
315 visual analogue scale, then transformed into a GAS scale depending on the pain level
316 considered as meeting the goal of a treatment; the goal for another client may be to

ACCEPTED MANUSCRIPT

317 decrease equinus (measured in degrees) in a gait analysis laboratory and the range of
318 motion is transformed into a GAS scale presenting a range of ankle positions as the goal.
319 However such a conversion is done at the expense of losing the linearity of the original
320 measure and although very useful in clinical practice, it is scientifically acceptable only if data
321 is analyzed as truly ordinal (therefore not using T-scores, nor means nor any arithmetical
322 operation) [65 66 28 67].

323 Can GAS levels be equidistant and GAS data be interval in form? This is possible if a
324 calibration by Rasch Analysis on an ”item bank” is carried out as proposed by Tennant et
325 al.[24], but at the expense of losing GAS adaptability to any goal [12]. Otherwise, GAS levels are
326 very unlikely to be interval despite all precautions, corrections and comparisons used to
327 minimize level inequality bias. Therefore ordinal interpretation of GAS, using rank tests (see
328 Steenbeek et al for an example[48]) and excluding all arithmetic procedures[66,68,24] on GAS
329 scores seems the most reasonable option for GAS data analysis. It is indisputably difficult to
330 set 5 equidistance GAS levels but simple rules can be postulated to facilitate choice of levels
331 (both for goal setters and the external judge who checks GAS levels and compares
332 equidistance between two groups): (1) avoid setting “half levels” (e.g. -0.5 as proposed by
333 Turner Stokes [69]); (2) have all clients start from the same initial score and therefore have all
334 clients assessed on the same number of levels of goal attainment (see[5] for a discussion on
335 advantages of scoring initial level at -2 or at -1).

336 What type of staff training should be required in order to use GAS in a research protocol?
337 Kiresuk, Smith and Cardillo had proposed that a minimum of one-year experience is required
338 to develop relevant and realistic scales[39]. Basic knowledge of GAS, and experience in
339 collaborative goal setting is not sufficient, as shown by Grant et al [62,12]. Reading a practical
340 guide[2,3,5,4] may be sufficient for GAS clinical use but not for studies aiming at producing valid

ACCEPTED MANUSCRIPT

341 and reliable GAS. A number of authors[70,41] propose practical training that is largely based on
342 formulating and correcting GAS scales, based on clients' real goals. We recommend that
343 such a practical training be used. This should also raise awareness on GAS quality criteria the
344 study will be judged on, and make GAS scoring easier after the intervention. The quality
345 appraisal here proposed could be the focus of such training.

346 Challenges in GAS methodology

347 GAS was intended to be a person-centered measure. Is there a risk of losing its person-
348 centered nature when trying to meet measurement criteria for controlled efficacy trials? The
349 risk of insisting upon an observable and measurable goal is to exclude family and client
350 appreciation of goal attainment. Appreciation of goal attainment by clients may be
351 subjective and is affected by a series of factors (self-awareness, denial, memory, high
352 involvement in goal pursuit) that can bias the perceived attainment of a goal; however goal
353 identification and measuring its attainment should be person-centered. The challenge for
354 clinicians and researchers is therefore to understand and analyze the client's (and/or his /her
355 families) goal and transform it into an observable, objective, performance-based measure
356 that can then be used to discuss goal attainment with the client. For example a general goal
357 of "improving my memory" expressed by a client can be transformed into a GAS scale that
358 measures memory functioning in real life situations agreed with the client (e.g. number of
359 medicines taken on time without prompting; number of failed-to-deliver messages in one
360 week…). Scoring GAS based on simple interview should be avoided when possible and
361 creativity used to link subjective goals in difficult domains to more objective goal attainment
362 indicators. For example for a GAS scale on anger management or use of social skills, rather
363 than asking the client to recall how often she/he felt he could cope with his anger or

ACCEPTED MANUSCRIPT

364 effectively use social/language skills, an effort should be made to choose indirect indicators
365 of goal attainment such as a review of controlled versus overt anger at the end of a day by
366 the client or proxy rating of social/language skill (how many times he/she
367 initiated/contributed to conversation or was understood) after regular naturally occurring
368 events (outing with friends, family dinner). It is probably the most challenging part of GAS
369 methodology but it has been demonstrated that it is feasible to have GAS that are both
370 person-centered in the choice of goal and objective in the formulation of GAS scales (see
371 Steenbeek et al. for an example[35]).

372 N-of-1 trial literature faces the same challenges and offers growing ingenuous methods
373 for assessment of domains not directly accessible to classical performance-based objective
374 measure. These methods include use of smartphone reminders to self-assess goal
375 attainment at regular periods to decrease memory bias (see[71] for an example), use of
376 naturally occurring situations monitored in real life by family/proxy to decrease self-
377 awareness bias (see[72] for an example), use of objective behavioral measures that are
378 thought to reflect the underlying psychological (e.g.: happiness[73]) or cognitive (e.g.:
379 functioning at school[74,75]) target goals. Future research should extend "goal menus" such as
380 those proposed by Turner-Stokes et al.[76] to more challenging domains (such as goals relating
381 to social and psychological functions, as well as goals focusing on performance in ecological
382 setting), using also N-of 1 trial literature. When goal attainment indicators are ecological and
383 monitored by the client or proxies, calculating an IRR is impossible (or difficult) and GAS
384 should be checked for items: precise description of all GAS levels, measurability,
385 unidimensionality and context of measurement to increase reliability in scoring. A
386 reasonable compromise between scientific rigor and person-centered approach could be to
387 have for each client at least one ecological client-centered GAS (with the risk of being less

ACCEPTED MANUSCRIPT

388 reliable) and at least one performance-based GAS (with the risk of being less ecological). In
389 all cases, Kiresuk recommended to "anchor scale points with behavioral or other evidence
390 that will be meaningful to the client and readily scored by the rater"[33] (p 31).

391 Limitations

392 The present article has a number of limitations. Choice of included criteria was not based
393 on a consensual agreement of all major teams using GAS in research but of four teams, from
394 three different countries. The aim of this paper was to raise awareness about the variability
395 of GAS use in published research and the need to build, in future, a consensus on the use of
396 GAS in efficacy research in rehabilitation. Although it may be viewed as a limitation, we
397 purposefully did not validate externally an appraisal score, so that the present guide acts
398 only as a starter for discussion and not a validated tool imposed on other teams that were
399 not included in the writing of the paper.
400

401 **Conclusion**

402 Goal Attainment Scaling has the potential to be sensitive to change following treatment
403 and applicable across divergent domains of rehabilitation, making it a useful rehabilitation
404 efficacy research outcome measure. However, GAS is used in studies with variable rigor that
405 impacts its validity and reliability, and therefore reduces the confidence one can have in the
406 results of a trial using GAS as an outcome measure. Clinimetric qualities of GAS are highly
407 dependent on the way GAS scales are written and therefore clinicians, researchers and
408 reviewers cannot rely on published studies of metrological qualities of GAS obtained in
409 different research studies. Researchers should be aware of the risk of bias related to the use
410 of imprecisely written GAS scales in research and make all possible efforts to minimize this

ACCEPTED MANUSCRIPT

411 bias by constructing high quality GAS scales for their clients, following recommendations
412 described here and previously published examples[27,48]. There is a need to develop in the
413 future a GAS quality appraisal score, similarly to standards used in other rehabilitation
414 controversial fields such as N-of-1 trials[51]. In this paper, we propose 17 criteria for appraising
415 GAS quality in trials using GAS as an outcome measure. We recommend that authors using
416 GAS report accurately how GAS methodology was used based on these criteria. The
417 reliability of these criteria needs to be established, but in the meantime we invite comment
418 and discussion of these proposals as we move towards a consensus on standards for use of
419 GAS in rehabilitation research.

420 **References**

421 1. Kiresuk TJ, Sherman R. Goal Attainment Scaling: A general method for evaluating
422 comprehensive community mental health programs. Comm Mental Health J.
423 1968;4:443–53.
424

425 2. Bovend'Eerdt TJ, Botell RE, Wade DT. Writing SMART rehabilitation goals and achieving
426 goal attainment scaling: a practical guide. Clin. Rehabil. 2009;23:352–61.
427

428 3. Turner-Stokes L. Goal Attainement Scaling: a practical guide.
429 http://www.bsrm.co.uk/eventdiaries/LeamingtonSpa2010/Docs/Goal%20Attainment%
430 20Scaling%20in%20Rehabilitation%20%20a%20practical%20guide.pdf.
431

432 4. Turner-Stokes L. Goal attainment scaling (GAS) in rehabilitation: a practical guide. Clin.
433 Rehabil. 2009;23:362–70.
434

435 5. Krasny-Pacini A, Hiebel J, Pauly F, Godon S, Chevignard M. Goal Attainment Scaling in
436 rehabilitation: A literature-based update. Ann. Phys. Rehabil. Med. 2013;56:212–30.
437

438 6. Schlosser RW. Goal attainment scaling as a clinical measurement technique in
439 communication disorders: a critical review. J. Commun. Disord. 2004;37:217–39.
440

ACCEPTED MANUSCRIPT

7. Hurn J, Kneebone I, Cropley M. Goal setting as an outcome measure: A systematic review. Clin. Rehabil. 2006;20:756–72.

8. Vu M, Law AV. Goal-attainment scaling: a review and applications to pharmacy practice. Res. Soc. Adm. Pharm. RSAP. 2012;8:102–21.

9. Malec JF. Goal Attainment Scaling in Rehabilitation. Neuropsychol. Rehabil. 1999;9:253–75.

10. Bouwens SFM, van Heugten CM, Verhey FRJ. The practical use of goal attainment scaling for people with acquired brain injury who receive cognitive rehabilitation. Clin. Rehabil. 2009;23:310–20.

11. Evans JJ. Goal setting during rehabilitation early and late after acquired brain injury. Curr. Opin. Neurol. 2012;25:651–5.

12. Grant M, Ponsford J. Goal attainment scaling in brain injury rehabilitation: strengths, limitations and recommendations for future applications. Neuropsychol. Rehabil. 2014;24:661–77.

13. Ostensjø S, Oien I, Fallang B. Goal-oriented rehabilitation of preschoolers with cerebral palsy--a multi-case study of combined use of the Canadian Occupational Performance Measure (COPM) and the Goal Attainment Scaling (GAS). Dev. Neurorehabilitation. 2008;11:252–9.

14. McDougall J, Wright V. The ICF-CY and Goal Attainment Scaling: Benefits of their combined use for pediatric practice. Disabil. Rehabil. 2009;31:1362–72.

15. Rushton PW, Miller WC. Goal attainment scaling in the rehabilitation of patients with lower-extremity amputations: A pilot study. Arch. Phys. Med. Rehabil. 2002;83:771–5.

16. Khan F, Pallant JF, Turner-Stokes L. Use of goal attainment scaling in inpatient rehabilitation for persons with multiple sclerosis. Arch. Phys. Med. Rehabil. 2008;89:652–9.

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

ACCEPTED MANUSCRIPT

disabling spinal disorders. Spine. 2009;34:2797–802.

18. Lewis VJ, Dell L, Matthews LR. Evaluating the feasibility of Goal Attainment Scaling as a rehabilitation outcome measure for veterans. J. Rehabil. Med. 2013;45:403–9.

19. Hanssen KT, Saltytė Benth J, Beiske AG, Landrø NI, Hessen E. Goal attainment in cognitive rehabilitation in MS patients. Neuropsychol. Rehabil. 2015;25:137–54.

20. 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.

21. Turner-Stokes L, Williams H, Johnson J. Goal attainment scaling: does it provide added value as a person-centred measure for evaluation of outcome in neurorehabilitation following acquired brain injury? J. Rehabil. Med. Off. J. UEMS Eur. Board Phys. Rehabil. Med. 2009;41:528–35.

22. Rockwood K, Joyce B, Stolee P. Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. J. Clin. Epidemiol. 1997;50:581–8.

23. Steenbeek D, Gorter JW, Ketelaar M, Galama K, Lindeman E. Responsiveness of Goal Attainment Scaling in comparison to two standardized measures in outcome evaluation of children with cerebral palsy. Clin. Rehabil. 2011;25:1128–39.

24. Tennant A. Goal attainment scaling: current methodological challenges. Disabil. Rehabil. 2007;29:1583–8.

25. 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.

26. Bovend'Eerdt TJH, Dawes H, Izadi H, Wade DT. Agreement between two different scoring procedures for goal attainment scaling is low. J. Rehabil. Med. Off. J. UEMS Eur. Board Phys. Rehabil. Med. 2011;43:46–9.

27. Ruble L, McGrew JH, Toland MD. Goal Attainment Scaling as an Outcome Measure in Randomized Controlled Trials of Psychosocial Interventions in Autism. J. Autism Dev. Disord. 2012;42:1974–83.

ACCEPTED MANUSCRIPT

28. MacKay G, Somerville W, Lundie J. Reflections on goal attainment scaling (GAS): cautionary notes and proposals for development. Educ. Res. 1996;38:161–72.

29. Turner-Stokes L, Baguley IJ, De Graaff S, Katrak P, Davies L, McCrory P, et al. Goal attainment scaling in the evaluation of treatment of upper limb spasticity with botulinum toxin: a secondary analysis from a double-blind placebo-controlled randomized clinical trial. J. Rehabil. Med. Off. J. UEMS Eur. Board Phys. Rehabil. Med. 2010;42:81–9.

30. Cytrynbaum S, Ginath Y, Birdwell J, Brandt L. Goal Attainment Scaling A Critical Review. Eval. Rev. 1979;3:5–40.

31. Simeonsson RJ, Jr DBB, Huntington GS, Brandon L. Scaling and attainment of goals in family-focused early intervention. Community Ment. Health J. 1991;27:77–83.

32. Ertzgaard P, Ward AB, Wissel J, Borg J. Practical considerations for goal attainment scaling during rehabilitation following acquired brain injury. J. Rehabil. Med. Off. J. UEMS Eur. Board Phys. Rehabil. Med. 2011;43:8–14.

33. Kiresuk TJ, Smith A, Cardillo JE, editors. Goal attainment scaling: Applications, theory, and measurement. Hillsdale, NJ, England: first published by Lawrence Erlbaum Associates, Inc, edition published in 2013 by Psychology Press, Hove, East Sussex, UK.; 1994.

34. Kiresuk TJ, Lund SH, Larsen NE. Measurement of goal attainment in clinical and health care programs. Drug Intell. Clin. Pharm. 1982;16:145–53.

35. 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.

36. P S, K R, Ra F, Dl S. The use of goal attainment scaling in a geriatric care setting. J. Am. Geriatr. Soc. 1992;40:574–8.

37. Palisano RJ. Validity of goal attainment scaling in infants with motor delays. Phys. Ther. 1993;73:651–8; discussion 658–60.

ACCEPTED MANUSCRIPT

38. King GA, McDougall J, Palisano RJ, Gritzan J, Tucker MA. Goal Attainment Scaling: its use in evaluating pediatric thearpy programs. Phys. Occup. Ther. Pediatr. 2000;19:31–52.

39. Smith A. Introduction and overview. In: Kiresuk TJ, Cardillo JE, editors. Goal attainment scaling: Applications, theory, and measurement. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; 1994. p. 1–14.

40. Stolee P, Stadnyk K, Myers AM, Rockwood K. An individualized approach to outcome measurement in geriatric rehabilitation. J. Gerontol. A. Biol. Sci. Med. Sci. 1999;54:M641–7.

41. 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.

42. Jones MC, Walley RM, Leech A, Paterson M, Common S, Metcalf C. Using goal attainment scaling to evaluate a needs-led exercise programme for people with severe and profound intellectual disabilities. J. Intellect. Disabil. JOID. 2006;10:317–35.

43. Mailloux Z, May-Benson TA, Summers CA, Miller LJ, Brett-Green B, Burke JP, et al. Goal attainment scaling as a measure of meaningful outcomes for children with sensory integration disorders. Am. J. Occup. Ther. Off. Publ. Am. Occup. Ther. Assoc. 2007;61:254–9.

44. Shefler 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.

45. 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.

46. Ottenbacher KJ, Cusick A. Discriminative Versus Evaluative Assessment: Some Observations on Goal Attainment Scaling. Am. J. Occup. Ther. 1993;47:349–54.

47. Zaza C, Stolee P, Prkachin K. The application of goal attainment scaling in chronic pain settings. J. Pain Symptom Manage. 1999;17:55–64.

48. 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.

ACCEPTED MANUSCRIPT

2005;19:274–82.

49. Sakzewski L, Boyd R, Ziviani J. Clinimetric properties of participation measures for 5- to 13-year-old children with cerebral palsy: a systematic review. Dev. Med. Child Neurol. 2007;49:232–40.

50. Ruble LA, Dalrymple NJ, McGrew JH. The Effects of Consultation on Individualized Education Program Outcomes for Young Children With Autism: The Collaborative Model for Promoting Competence and Success. J. Early Interv. 2010;32:286–301.

51. Tate RL, Perdices M, Rosenkoetter U, Wakim D, Godbee K, Togher L, et al. Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: the 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. Neuropsychol. Rehabil. 2013;23:619–38.

52. Bergsma D, Baars-Elsinga A, Sibbel J, Lubbers P, Visser-Meily A. Visual daily functioning of chronic stroke patients assessed by goal attainment scaling after visual restorative training: an explorative study. Top. Stroke Rehabil. 2014;21:400–12.

53. Peirone E, Goria PF, Anselmino A. A dual-task home-based rehabilitation programme for improving balance control in patients with acquired brain injury: a single-blind, randomized controlled pilot study. Clin. Rehabil. 2014;28:329–38.

54. Phillips MF, Robertson Z, Killen B, White B. A pilot study of a crossover trial with randomized use of ankle-foot orthoses for people with Charcot-Marie-tooth disease. Clin. Rehabil. 2012;26:534–44.

55. Wallen MA, O'Flaherty SJ, Waugh M-CA. Functional outcomes of intramuscular botulinum toxin type a in the upper limbs of children with cerebral palsy: a phase II trial. Arch. Phys. Med. Rehabil. 2004;85:192–200.

56. Bjornson K, Hays R, Graubert C, Price R, Won F, McLaughlin JF, et al. Botulinum toxin for spasticity in children with cerebral palsy: a comprehensive evaluation. Pediatrics. 2007;120:49–58.

57. Lowe K, Novak I, Cusick A. Repeat injection of botulinum toxin A is safe and effective for upper limb movement and function in children with cerebral palsy. Dev. Med. Child Neurol. 2007;49:823–9.

ACCEPTED MANUSCRIPT

58. Bovend'Eerdt TJ, Dawes H, Sackley C, Izadi H, Wade DT. An integrated motor imagery program to improve functional task performance in neurorehabilitation: a single-blind randomized controlled trial. Arch. Phys. Med. Rehabil. 2010;91:939–46.

59. Berge SR Ten, Boonstra AM, Dijkstra PU, Hadders-Algra M, Haga N, Maathuis CGB. A systematic evaluation of the effect of thumb opponens splints on hand function in children with unilateral spastic cerebral palsy. Clin. Rehabil. 2012;26:362–71.

60. Leach E, Cornwell P, Fleming J, Haines T. Patient centered goal-setting in a subacute rehabilitation setting. Disabil. Rehabil. 2009;32:159–72.

61. Malec JF, Smigielski JS, DePompolo RW. Goal attainment scaling and outcome measurement in postacute brain injury rehabilitation. Arch Phys Med Rehabil. 1991;72:138–43.

62. Grant M, Ponsford J, Bennett PC. The application of Goal Management Training to aspects of financial management in individuals with traumatic brain injury. Neuropsychol. Rehabil. 2012;22:852–73.

63. Krasny-Pacini A, Limond J, Evans J, Hiebel J, Bendjelida K, Chevignard M. Context-Sensitive Goal Management Training for Everyday Executive Dysfunction in Children After Severe Traumatic Brain Injury. J. Head Trauma Rehabil. 2014;

64. Mall V, Heinen F, Siebel A, Bertram C, Hafkemeyer U, Wissel J, 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.

65. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch. Phys. Med. Rehabil. 1989;70:308–12.

66. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. J. Clin. Epidemiol. 1996;49:711–7.

67. Grenville J, Lyne P. Patient-centred evaluation and rehabilitative care. J. Adv. Nurs. 1995;22:965–72.

68. Küçükdeveci AA, Tennant A, Grimby G, Franchignoni F. Strategies for assessment and outcome measurement in physical and rehabilitation medicine: an educational review.

ACCEPTED MANUSCRIPT

J. Rehabil. Med. Off. J. UEMS Eur. Board Phys. Rehabil. Med. 2011;43:661–72.

69. Turner-Stokes L, Williams H. Goal attainment scaling: a direct comparison of alternative rating methods. Clin. Rehabil. 2010;24:66–73.

70. Cardillo JE. Goal Setting, Follow-up and Goal Monitoring. In: Kiresuk TJ, Smith A, editors. Goal attainment scaling: Applications, theory, and measurement. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; 1994. p. 39–59.

71. Arco L. Neurobehavioural treatment for obsessive-compulsive disorder in an adult with traumatic brain injury. Neuropsychol. Rehabil. 2008;18:109–24.

72. Baldwin VN, Powell T. Google Calendar: A single case experimental design study of a man with severe memory problems. Neuropsychol. Rehabil. 2014;1–20.

73. Logan KR, Jacobs HA, Gast DL, Murray AS, Daino K, Skala C. The Impact of Typical Peers on the Perceived Happiness of Students with Profound Multiple Disabilities. Res. Pract. Pers. Sev. Disabil. 1998;23:309–18.

74. Feeney TJ, Ylvisaker M. Context-sensitive behavioral supports for young children with TBI: short-term effects and long-term outcome. J. Head Trauma Rehabil. 2003;18:33–51.

75. Feeney T, Ylvisaker M. Context-sensitive cognitive-behavioural supports for young children with TBI: A replication study. Brain Inj. 2006;20:629–45.

76. Ashford S, Jackson D, Turner-Stokes L. Goal setting, using goal attainment scaling, as a method to identify patient selected items for measuring arm function. Physiotherapy. 2015;101:88–94.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Quality appraisal items	Item description	Examples of reported criteria, extracted from rehabilitation effectiveness studies and/or published methods that satisfy criteria.	Potential for bias arising from failure to report criteria and/or utility of reporting the criterion.
Content Validity			
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"* [1] *"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."*[2].	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[3] 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 [4] 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 [5]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. [6] 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	In an intervention for executive dysfunction in children [7], 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).

ACCEPTED MANUSCRIPT

	intervention is aiming to improve.	intervention, but they might have been used as "control" goals, not expected to be attained.	
Reliability			
Reliability of scale construction			
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 "*[7].	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)".* [8 9] In a study of infants with motor delays, Palisano[3] 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[4] reported that realism of the expected level of outcome	An experimental group may falsely present higher GAS scores post intervention because GAS scales were formulated with easier levels of goal attainment.

ACCEPTED MANUSCRIPT

		"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.	
Time-specificity	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.	In the RCT of Lowe et al[1], 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.	Goal difficulty across experimental groups may have been unequal, at a given assessment point, due to differences in time-frames for goal completion.
Reliability of scale rating			
Inter-rater reliability (IRR)	Inter-rater reliability of GAS scales is reported.	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 [10] .	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.
Criteria affecting IRR			
Precise description of all levels	Five GAS levels have been precisely described pre-intervention for each scale.	*"Goal: Reducing weekly shopping expenditure* *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"* [11]	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.
Measurability	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.	-good measurability: 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 […]*[12]*)."* -unclear measurability: 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[13]). See Ruble et al. for an example of assessment of measurability of	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.

ACCEPTED MANUSCRIPT

		social and cognitive goals[8 9].	
Unidimensionality	GAS scales have been verified by an external judge to check for unidimensionality.	-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 […]"*[13] -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]	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.
Context of measurement	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].	e.g.: setting/prompting/guidance: *"in order to create an irregular surface, a ladder is placed horizontally at a height of 15cm, the girl is asked to walk barefoot without orthosis, as quickly as possible, through the rungs over a distance of 8 meters. Only if she falls a therapist will help her holding one of her hands."* [17] e.g.: *"GAS level -1: Prepares school bag but requires constant verbal guidance from the parents or teacher; GAS level 0: Manages to prepare the school bag using a check-list of necessary steps and under supervision; GAS level+1: Manages to prepare school bag alone, using a check-list of necessary steps; GAS level +2: No supervision required, child only occasionally forgets items."*[14 7]	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.
Other criteria			
Training	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.	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]).	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…).
Examiner bias	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).	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]. e.g.:"The therapist-GAS was scored from video by blind evaluators. The parent-GAS was scored by two blinded occupational therapists."[1].	If the same person sets the GAS and scores them, he/she is likely to be biased towards scoring a maximum of 0 ("attained as expected"). He/she may rely on memory of initial performance and subjective impression of improvement to score ambiguous progress. The independence of assessor should also be

ACCEPTED MANUSCRIPT

			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).	*e.g.:"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"*[10].	The performance of arithmetic operations such as T-scores on ordinal data is scientifically not valid[23] 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 [10 11] [24] in the methods or results section. Examples of goal types can be given (1) by providing examples of goal in the paper[25 26 27 2 28]; (2) by reporting all chosen goals in appendix[25 11]; (3) by reporting goal type and frequency of each type without providing an exhaustive goal list [22 29 10].	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. [9] 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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

1. Lowe K, Novak I, Cusick A. Repeat injection of botulinum toxin A is safe and effective for upper limb movement and function in children with cerebral palsy. Dev. Med. Child Neurol. 2007;49:823–9.

2. Bjornson K, Hays R, Graubert C, Price R, Won F, McLaughlin JF, et al. Botulinum toxin for spasticity in children with cerebral palsy: a comprehensive evaluation. Pediatrics. 2007;120:49–58.

3. Palisano RJ. Validity of goal attainment scaling in infants with motor delays. Phys. Ther. 1993;73:651–8; discussion 658–60.

4. Cardillo JE. Goal Setting, Follow-up and Goal Monitoring. In: Kiresuk TJ, Smith A, editors. Goal attainment scaling: Applications, theory, and measurement. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; 1994. p. 39–59.

5. Turner-Stokes L, Baguley IJ, De Graaff S, Katrak P, Davies L, McCrory P, et al. Goal attainment scaling in the evaluation of treatment of upper limb spasticity with botulinum toxin: a secondary analysis from a double-blind placebo-controlled randomized clinical trial. J. Rehabil. Med. Off. J. UEMS Eur. Board Phys. Rehabil. Med. 2010;42:81–9.

6. Phillips MF, Robertson Z, Killen B, White B. A pilot study of a crossover trial with randomized use of ankle-foot orthoses for people with Charcot-Marie-tooth disease. Clin. Rehabil. 2012;26:534–44.

7. Krasny-Pacini A, Limond J, Evans J, Hiebel J, Bendjelida K, Chevignard M. Context-Sensitive Goal Management Training for Everyday Executive Dysfunction in Children After Severe Traumatic Brain Injury. J. Head Trauma Rehabil. 2014;

8. Ruble LA, Dalrymple NJ, McGrew JH. The Effects of Consultation on Individualized Education Program Outcomes for Young Children With Autism: The Collaborative Model for Promoting Competence and Success. J. Early Interv. 2010;32:286–301.

9. Ruble L, McGrew JH, Toland MD. Goal Attainment Scaling as an Outcome Measure in Randomized Controlled Trials of Psychosocial Interventions in Autism. J. Autism Dev. Disord. 2012;42:1974–83.

10. 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.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

11. Grant M, Ponsford J, Bennett PC. The application of Goal Management Training to aspects of financial management in individuals with traumatic brain injury. Neuropsychol. Rehabil. 2012;22:852–73.

12. 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.

13. Malec JF. Goal Attainment Scaling in Rehabilitation. Neuropsychol. Rehabil. 1999;9:253–75.

14. Krasny-Pacini A, Hiebel J, Pauly F, Godon S, Chevignard M. Goal Attainment Scaling in rehabilitation: A literature-based update. Ann. Phys. Rehabil. Med. 2013;56:212–30.

15. Grant M, Ponsford J. Goal attainment scaling in brain injury rehabilitation: strengths, limitations and recommendations for future applications. Neuropsychol. Rehabil. 2014;24:661–77.

16. Bovend'Eerdt TJ, Botell RE, Wade DT. Writing SMART rehabilitation goals and achieving goal attainment scaling: a practical guide. Clin. Rehabil. 2009;23:352–61.

17. 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.

18. Mailloux Z, May-Benson TA, Summers CA, Miller LJ, Brett-Green B, Burke JP, et al. Goal attainment scaling as a measure of meaningful outcomes for children with sensory integration disorders. Am. J. Occup. Ther. Off. Publ. Am. Occup. Ther. Assoc. 2007;61:254–9.

19. Turner-Stokes L. Goal attainment scaling (GAS) in rehabilitation: a practical guide. Clin. Rehabil. 2009;23:362–70.

20. Bovend'Eerdt TJ, Dawes H, Sackley C, Izadi H, Wade DT. An integrated motor imagery program to improve functional task performance in neurorehabilitation: a single-blind randomized controlled trial. Arch. Phys. Med. Rehabil. 2010;91:939–46.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

21. Mendelson AD, McCullough C, Chan A. Integrating self-management and exercise for people living with arthritis. Health Educ. Res. 2011;26:167–77.

22. Wallen MA, O'Flaherty SJ, Waugh M-CA. Functional outcomes of intramuscular botulinum toxin type a in the upper limbs of children with cerebral palsy: a phase II trial. Arch. Phys. Med. Rehabil. 2004;85:192–200.

23. Tennant A. Goal attainment scaling: current methodological challenges. Disabil. Rehabil. 2007;29:1583–8.

24. Dahlberg CA, Cusick CP, Hawley LA, Newman JK, Morey CE, Harrison-Felix CL, et al. Treatment efficacy of social communication skills training after traumatic brain injury: a randomized treatment and deferred treatment controlled trial. Arch. Phys. Med. Rehabil. 2007;88:1561–73.

25. Bergsma D, Baars-Elsinga A, Sibbel J, Lubbers P, Visser-Meily A. Visual daily functioning of chronic stroke patients assessed by goal attainment scaling after visual restorative training: an explorative study. Top. Stroke Rehabil. 2014;21:400–12.

26. Peirone E, Goria PF, Anselmino A. A dual-task home-based rehabilitation programme for improving balance control in patients with acquired brain injury: a single-blind, randomized controlled pilot study. Clin. Rehabil. 2014;28:329–38.

27. Berge SR Ten, Boonstra AM, Dijkstra PU, Hadders-Algra M, Haga N, Maathuis CGB. A systematic evaluation of the effect of thumb opponens splints on hand function in children with unilateral spastic cerebral palsy. Clin. Rehabil. 2012;26:362–71.

28. Jones MC, Walley RM, Leech A, Paterson M, Common S, Metcalf C. Using goal attainment scaling to evaluate a needs-led exercise programme for people with severe and profound intellectual disabilities. J. Intellect. Disabil. JOID. 2006;10:317–35.

29. Wallen M, O'Flaherty SJ, Waugh M-CA. Functional Outcomes of Intramuscular Botulinum Toxin Type A and Occupational Therapy in the Upper Limbs of Children With Cerebral Palsy: A Randomized Controlled Trial. Arch. Phys. Med. Rehabil. 2007;88:1–10.

ACCEPTED MANUSCRIPT