TPM Vol. 33, No. 1, 2026
ISSN: 1972-6325
https://www.tpmap.org/

INTER-USER RELIABILITY AND TEMPORAL STABILITY OF CONVENTIONAL AND MODIFIED GOAL ATTAINMENT SCALING SCALE CONSTRUCTION IN NOVICE USERS

JAMIE BERRY
MACQUARIE UNIVERSITY

SAGAR NAGARAJ
ADVANCED NEUROPSYCHOLOGICAL TREATMENT SERVICES

JENNIFER KHOUW
ADVANCED NEUROPSYCHOLOGICAL TREATMENT SERVICES

E. ARTHUR SHORES
MACQUARIE UNIVERSITY

JENNIFER BATCHELOR
MACQUARIE UNIVERSITY

Abstract:
For over 50 years, Goal Attainment Scaling (GAS) has been a useful but controversial tool in evaluating outcomes for individuals undergoing rehabilitation. Despite measuring meaningful goals, it has been criticized in the literature for its unstandardized and imprecise GAS scale construction that results in questionable reliability and validity. The aim of the current study was to assess and compare the inter-user reliability and temporal stability of the construction of GAS scales using a conventional (Flexible) and modified (Structured) approach. First-year university psychology students (N = 79) constructed GAS scales for six separate goal-setting scenarios based on the GAS condition (Structured or Flexible) they were randomly assigned to, and they were invited to repeat GAS scale construction for the same scenarios two weeks later. GAS scale median and range scores were compared across raters within each group at Time 1 to assess inter-user reliability and across Time 1 and Time 2 to assess temporal stability. Structured GAS demonstrated higher inter-user reliability than Flexible GAS for both the median ($\kappa = 0.28$; $\kappa = 0.19$) and range scores ($\kappa = 0.30$; $\kappa = 0.05$). Structured GAS also demonstrated higher temporal stability across a 2-week period, with greater overall Kappa values for the median and range scores.
Conclusions: Structured GAS demonstrated greater inter-user reliability and temporal stability than Flexible GAS among novice GAS users.
Keywords: Goal attainment scaling, Reliability, Validity, Rehabilitation, Goal setting

INTRODUCTION:

Goal Attainment Scaling (GAS) is a method of measuring individually tailored goal outcomes, originally designed for use in outpatient mental health services, and broadly adopted in rehabilitation settings (Cheema et al., 2024; Malec, 1999; Krasny-Pacini et al., 2013). The methodology of GAS can be summarized as a six-step process (Stolee et al., 1992; Malec, 1999). The first step involves identifying the person's problem areas and refining a group of priority goals. These selected goals are then weighted by their importance to the person and a follow up period is established. A five-point Goal Attainment Scale is constructed, with the most likely outcome of intervention set at the midpoint of the scale or the '0' level, the most favorable at +2 and the least favorable at -2, with values representing partially better or worse goal attainment set at the +1 and -1 levels respectively (Kiresuk & Sherman, 1968). Upon assessment at the predetermined follow-up time point, GAS scores can be converted into a standardised T-score, which purportedly enables the comparison of GAS scores across different people and goals (Kiresuk & Sherman, 1968), although use of GAS T-scores has been criticized due to the data being ordinal and not equidistant across the GAS levels (Krasny-Pacini et al., 2016).

Reliability of GAS

The factors influencing the reliability and validity of GAS as a measurement tool generally pertain to therapist bias and error in developing specific and measurable intervention goals, as well as setting appropriate values for each level of the GAS scale (Ruble et al., 2012). For example, therapists may set goals that are clinically non-

meaningful or too easy to achieve or construct scales that are ambiguous to score due to multidimensionality, limited range, gaps, overlap or non-exclusivity of GAS levels (Grant & Ponsford, 2014). The reliability of GAS has been evaluated primarily in terms of its inter-rater and test-retest reliability (Kiresuk, 1973; Hurn, Kneebone, & Cropley, 2006; Steenbeek et al., 2010; Krasny-Pacini et al., 2013). In a review of the validity of GAS scales, many studies included data on inter-rater reliability (73.0% of reviewed studies), test–retest reliability (16.2% of reviewed studies) and internal consistency (13.5% of reviewed studies; Shankar et al., 2020).

Inter-rater Reliability

Inter-rater reliability has generally been assessed by measuring the agreement between post- intervention GAS ratings on the same GAS scales (Hurn et al., 2006; Steenbeek et al., 2010). The literature has generally described the inter-rater reliability of GAS ratings as adequate (Krasny-Pacini et al., 2013), with the agreement between judges' scores ranging from moderate (κ=.48; Bovend'Eerdt et al., 2011) to excellent (κ=.91; Rockwood et al., 1997). The inter-rater reliability of GAS ratings can be heavily influenced by the rater's experience with GAS and the person being rated, as this affects both how scales are constructed and scored (Cytrynbaum et al., 1979; Steenbeek et al., 2010; Krasny-Pacini et al., 2016). Steenbeek et al. (2010) demonstrated that on average, the inter-rater reliability of GAS ratings on scales constructed by the therapists treating the person (κ=.82) was greater than for ratings on scales constructed by independent therapists (κ=.64), with the main reason for disagreement being the discrepancies in therapists' interpretations of client behaviors. Independent therapists have less nuanced knowledge of specific patient behaviors or challenges than treating therapists, and are therefore less able to extract relevant goal setting information in one session, particularly if the patient has difficulties effectively communicating their situation (Bovend'Eerdt et al., 2011). In such cases, Bovend'Eerdt et al. (2011) found the inter-rater reliability between the treating therapist and the independent therapist to be low, with an intraclass correlation of 0.48 that held true above and beyond therapist training with GAS and experience with a cognitively challenged population. Thus, overall, the inter-rater reliability of GAS ratings appears to be moderate to excellent (Landis & Koch, 1977), although this varies based on the extent of collaboration between the patient and therapist, the experience of the therapist, and the construction of the GAS scale (Steenbeek et al., 2010; Bovend'Eerdt et al., 2011; Krasny-Pacnini et al., 2013).

Test-retest Reliability

Comparatively, the test-retest reliability of GAS has been investigated far less in the GAS literature (Schlosser 2004; Steenbeek et al., 2007). In their review, Hurn et al. (2006) identified only one study that measured the test-retest reliability of GAS (Cornbleth, 1978). However, Kiresuk (1973) reported test- retest reliabilities of r = .70 for GAS outcome scores and r = .88 for GAS scale content on follow-up scales scored by two teams of raters. In a sample of twenty occupational therapy students, Koski and Richards (2015) reported very strong test-retest reliability for GAS ratings at three weeks and one day compared with three weeks and three days after goal setting, ranging from r = .749 to r = 1.00. The findings however were limited by the small sample size and the short retest period of only two days (Koski & Richards, 2015). The test-retest reliability of GAS may be directly influenced by the experience and training of staff (Steenbeek et al., 2007), however, research in this area of GAS is lacking (Cox & Amsters, 2002). According to Cytrynbaum et al. (1979, p.26), test-retest reliability was not viable for a non-standardised measure like GAS, as there was no way to avoid the "person variance being confounded with the rater variance" when assessing the same person rating two goals at two different time periods.

Inter-user Reliability of Scale Construction

The Krasny-Pacini et al. (2016) reliability of scale rating criteria include criteria for inter-rater reliability, as well as the following four criteria proposed to affect inter-rater reliability: precise description of all levels, measurability, unidimensionality, and context of measurement. The reliability of scale construction criteria refer to the following scale construction considerations: equidistance of levels, preintervention performance, attainability/difficulty, and time-specificity. However, none of these criteria refer to how likely two GAS scale constructors are to construct the same GAS scale. Such a measure would go some way in establishing consistency of GAS scale construction when all other goal setting data are held constant, such as the qualitative description of the goal or goal area and the person for whom the goal is being set. Whilst this type of reliability is akin to the concept of inter-rater reliability, that term is misleading because it typically relates to the reliability of the GAS rating, rather than scale construction. We propose that the term inter-user reliability is more fitting to describe this important consideration that is lacking in the GAS literature. That is, an examination of the extent to which two or more independent goal setters will construct the same GAS scale, holding all other variables constant.

Some researchers have investigated aspects of inter-user reliability. For example, May-Benson et al. (2021) found there was 78% agreement in GAS goal content among therapists constructing goals for children with sensory processing disorder based on parent interviews. With a similar focus on qualitative goal content, Rushton et al. (2002) found 63% agreement of goals between independent investigators setting goals with patients with lower-extremity amputations in a rehabilitation setting. However, in each of these studies, the focus was on the qualitative goal area, rather than the numerical variables pertaining to the target goal (i.e., range of values defining the GAS scale broadly and/or the expected outcome specifically).

Temporal Stability of Scale Construction

The Krasny-Pacini et al (2016) guidelines do not include any criteria that address the consistency of scale construction across time, which is similar to test-retest reliability, albeit with a focus on scale construction, rather than GAS rating. In other words, given a certain goal setting scenario, how likely is the same GAS scale constructor to construct the same GAS scale at a different point in time? The problem of person variance being confounded with rater variance (Cytrynbaum et al., 1979) applies to examining test-retest reliability of GAS ratings, but not GAS scale construction. We have identified no literature examining the reliability of GAS scale construction at two time points, which we term temporal stability of scale construction. One approach to this problem is to hold the 'person variance' constant by having participants construct GAS scales based on controlled goal setting scenarios at two time-points, which was adopted in the current study.

Current Study

Based on some of the challenges associated with GAS scale construction (Ruble et al., 2012; Krasny- Pacini et al., 2016), a modified approach to GAS was devised by Berry et al. (2023) that utilizes a purpose-built calculator to produce the GAS scale (Clark et al., 2021; neurotreatment.com.au/goal-attainment-scaling-range-generator.aspx). The calculator algorithm ensures that after entering the person's baseline performance (at the mid-point of the -2 GAS level) and the maximum realistic performance (ceiling of the +2 GAS level), a five-point scale with continuous and attainable values are produced (Berry et al., 2023). In the current study, this modified GAS method is termed 'Structured GAS', and conventional GAS is termed 'Flexible GAS', reflecting its free-hand approach. The aim of this research was to assess and compare the inter-user reliability and temporal stability of Structured and Flexible GAS scale construction in novice users. The modified approach to GAS also incorporates use of a goal menu and control goals (Berry et al., 2023), however, those elements were not applied in the current study. Based on the standardized approach of Structured GAS, it was hypothesized that it would show higher inter-user reliability and temporal stability of scale construction than Flexible GAS.

METHOD

Participants

First year psychology students (N = 79) from [removed for peer review] University were recruited via the Department of Psychology Participant Pool (SONA) website. Participants were excluded from the study if they were under 18 or above 70 years of age, had less than intermediate proficiency in English, did not hold at least a school certificate, or had previous experience with GAS or other goal setting therapy from a registered clinician. This ensured that any major confounding factors affecting GAS scale construction were mitigated and that the sample represented novice users of GAS (Steenbeek et al., 2007).

Measures

Flexible GAS.

Flexible GAS scales were constructed using the procedure outlined in Table 1. T-scores were not calculated due to their inappropriate use given the data are non-parametric (Krasny-Pacini et al., 2016). Given weightings are incorporated into the T-score calculation, this step was also omitted.

TABLE 1 GAS Procedures

Flexible GAS	Structured GAS
1. Goal selection	1. Goal selection
2. Set time period for goal attainment	2. Set time period for goal attainment
3. Define expected level of goal attainment (0)	3. Define Current Level of Functioning
4. Define all other levels of goal attainment (-2, -1, +1, +2)	4. Define Maximum Realistic Level of Functioning
	5. Generate GAS scale via online calculator

Note. GAS = Goal Attainment Scaling

Structured GAS.

In Structured GAS (Table 1), the first two steps of the process are identical to Flexible GAS, however, the rater is asked to identify the current level of functioning (CLF) and maximum realistic level of functioning (MRLF) of the goal behaviour and enter these values into the GAS calculator in order to generate the values for the five GAS levels. The CLF is the current state of the individual's performance on the goal behaviour and the MRLF is the maximum or best possible performance of the behaviour that is attainable within the defined time period (Berry et al., 2023). For example, if the goal was to eat two out of three meals per day, then the MRLF would be to eat 3 meals per day or 21 meals per week. Although the CLF may be set at the -1 level when there is a chance the person could deteriorate in functioning in the goal area (Turner-Stokes, 2009; Grant & Ponsford, 2014), setting the baseline at the -1 level limits the range of improvement from 4 levels to 3 and is therefore not

recommended, especially when decline is unlikely (Ruble et al., 2012). Therefore, the calculator sets the CLF at the midpoint of the -2 level. The MRLF was set at the upper limit of the +2 level as this was the ceiling of the scale. With both the CLF and MRLF values, the calculator used a formula to generate five roughly equal sized range values with no overlap (Berry et al., 2023; Clark et al., 2021).

Goal setting scenarios.

The six goal setting scenarios were developed by a senior clinical neuropsychologist. They were developed with the intention of i) being relatable to naive participants with no experience in rehabilitation, ii) being distinct from one another, iii) containing sufficient information to set a GAS scale, and iv) tapping into only one type of goal behaviour. For each scenario, participants were required to define the GAS values for each level of the GAS scale and determine the frequency (per day, per week, per fortnight, per month) of the goal behaviour. This approach maximized the chances of participants providing numerical values that could then be analyzed quantitatively.

Procedure

Participants were randomly assigned to the Structured or Flexible GAS condition and presented with the corresponding instructions on how to set and scale goals using the particular GAS type. These instructions were followed by an associated example and practice scenario that required participants to fill out a GAS scale using the designated type of GAS. Participants were provided with feedback on a valid way to construct the GAS scale for the practice example before they were asked to construct GAS scales for six goal setting scenarios. Following completion, participants were asked to construct GAS scales based on the same six scenarios two weeks later, if they had opted to complete a second session. The study was approved by the Macquarie University Human Research Ethics Committee (reference number: 52020624915310).

Statistical Analysis

Statistical procedures were carried out using StataCorp Stata Version 16 and JASP v0.17.3.

Data Standardisation and Adjustment.

To allow statistical comparison, all data provided by participants were standardised to the same time period (month). For example, a goal behaviour of "Wake up early for work five times per week" or "Wake up early for work every day" were both standardized to "Wake up early for work 20 times per month". If the participant put in a range of values for a level, the average of the values was standardized to a month. For example, "Wake up early four to six times per week" was standardized to "Wake up early 20 times per month".

Some data were adjusted when participants made small but obviously incorrect errors in setting the time period, e.g. 'Lucy arrives to work on time five times per day' was corrected to 'five times per week'. Some data were also reverse scored, as a large portion of participants had represented the +2 level as the absence of a negative outcome. For example, in the 'Lucy arriving to work' scenario, some participants wrote '+2: Arrive at work late no times per week', while other participants wrote '+2: Arrive at work on time five times per week'. As these were equivalent, these data were standardized to the same value, 'arrive on time 20 times per month'. No further adjustments were made to the data.

Hypothesis 1: Inter-user Reliability.

Median and range values are proposed to be defining properties of quantified univariate GAS scales. The median represents the value corresponding to the expected outcome and the range defines the numerical scope corresponding to all possible outcomes for a univariate GAS scale. Inter-user reliability was established with a Fleiss' Kappa analysis, measuring the proportion of agreement between median and range scores for Structured and Flexible GAS at Time 1 only. Unlike Cohen's Kappa which is restricted to measuring the agreement between two raters, Fleiss' Kappa can be generalized to cases with more than two raters scoring the same set of scenarios (Fleiss, 1971; Fleiss, Levin, & Paik, 1981). It is important to note that the value of the Kappa statistic does not represent the correlation between scores, but rather the proportion of agreement between participants that is greater than chance (Fleiss, 1971; Sim & Wright, 2005). Therefore, a significant value indicates that agreement between participants is due to the scale and different to chance (Sim & Wright, 2005). The Landis & Koch (1977) benchmark strengths of agreement were applied, being <0.00 – poor; 0.00-0.20 – slight; 0.21-0.40 – fair; 0.41-0.60 – moderate; 0.61-0.80 – substantial; 0.81-1.00 – almost perfect. Additionally, Z-tests were used to compare the Kappa coefficients across conditions.

Hypothesis 2: Temporal Stability.

Given temporal stability involved the same rater at two time points, Cohen's Kappa was used to measure the level of agreement of each participant's median and range scores across Time 1 and Time 2, for both the Flexible and Structured GAS conditions. Z-tests were used to compare the Kappa coefficients across conditions.

RESULTS

There were n = 33 participants in the Structured GAS and n = 46 participants in the Flexible GAS conditions at Time 1; the uneven numbers in groups was caused by an unintended error in the allocation process whereby the

randomization procedure was set to an unequal ratio. Table 2 details the demographic characteristics of participants in the Flexible and Structured GAS groups for Time 1 and Time 2. In general, most participants were female and had completed their secondary education.

TABLE 2 Sample Characteristics

	Time 1				Time 2			
	Flexible GAS (n=46)		Structured GAS (n=33)		Flexible GAS (n=13)		Structured GAS (n=13)	
	M (SD)	Range	M (SD)	Range	M (SD)	Range	M (SD)	Range
Age	20.0 (3.2)	18 - 52	23.2 (8.3)	18 - 47	24.8 (11.7)	18 - 52	21.7 (5.0)	18 - 34
	n (%)		n (%)		n (%)		n (%)	
Gender								
Male	13 (29)		6 (18)		5 (38)		3 (23)	
Female	33 (71)		27 (82)		8 (62)		10 (77)	
Education								
Tertiary	7 (15)		5 (15)		4 (31)		2 (16)	
Secondary	34 (74)		24 (73)		7 (54)		10 (77)	
Diploma	3 (7)		2 (6)		2 (16)			
Trade	1 (2)		0 (0)					
Year 10	1 (2)		2 (6)				1 (8)	

Note. GAS = Goal Attainment Scaling.

Hypothesis 1: Inter-user Reliability.
Table 3 describes the levels of agreement between the median and range scores across participants within each GAS condition. Across participants in Structured GAS, there was fair agreement between the median and range scores (Table 4). In Flexible GAS there was slight agreement of median and range scores between participants. The kappa coefficient for Structured GAS median scores was higher than that for Flexible GAS ($Z = 12.6$, $p < .001$) and the kappa coefficient for Structured GAS range scores was higher than that for Flexible GAS ($Z = 38.9$, $p < .001$).

TABLE 3. Combined Kappa Values for Median and Range Scores across GAS Condition (95% confidence intervals) Assessing Inter-user Reliability

Condition	Median	Range
Structured GAS	0.28* (.27 - .29)	0.30* (.29 - .31)
Flexible GAS	0.19* (.18 - .20)	0.05* (.05 - .06)

Note. GAS = Goal Attainment Scaling.
*p <.01

Hypothesis 2: Temporal Stability.
Thirty-three percent of participants elected to return for a second session two weeks after the first. Table 2 shows the characteristics of the retest sample, which contained n=13 participants in each group. Table 4 shows that in Structured GAS, there was fair to moderate temporal stability of median scores in all goal-setting scenarios except scenario 2. Similarly in Flexible GAS, temporal stability of median scores was between the fair and moderate range, with the exception of scenarios 1 and 5, which were not significantly different to chance.

TABLE 4. Kappa Values [and Observed, Expected Agreement Percentages] for Median and Range Values Assessing Temporal Stability

		Scenario					
		1	2	3	4	5	6
Medians							
	Structured GAS	0.43**	-0.08[a]	0.47**	0.31**	0.59**	0.43**
		[84.62,	[69.23,	[61.54,	[46.15,	[76.92,	[76.92,
		72.78]	71.60]	27.22]	21.89]	43.79]	59.76]
	Flexible GAS	0.13[a]	0.52**	0.59**	0.30**	0.12	0.27*
		[61.54,	[69.23,	[69.23,	[38.46,	[30.77,	[46.15,
		55.62]	36.09]	24.26]	11.83]	21.30]	26.04]
Ranges							
	Structured GAS	0.54**	-0.08[a]	0.25**	0.23*	0.41**	0.20*
		[84.62,	[76.92,	[30.77,	[38.46,	[69.23,	[76.92,
		66.27]	78.70]	8.28]	20.12]	47.93]	71.01]
	Flexible GAS	-0.02	0.06	0.27**	0.02	-0.02	0.09
		[7.69,	[15.38,	[30.77,	[7.69,	[0.00,	[15.38,
		9.47]	10.06]	5.33]	5.92]	1.78]	6.51]

Note. GAS = Goal Attainment Scaling.
[a] Paradox
*p<.05, **p<.01

The temporal stability of range scores in Structured GAS was mostly fair to moderate, with only scenario 6 eliciting slight agreement and scenario 2 not reaching significance. In Flexible GAS, only scenario 3 had fair temporal stability, while the agreement between range scores on all other scenarios were not significantly different to chance (Table 4). The kappa coefficients for Structured GAS median scores were higher than those for Flexible GAS for scenarios 1 ($Z = 4.62$, $p < .001$) and 5 ($Z = 3.9$, $p < .001$), whereas the kappa coefficient was higher for Flexible GAS for scenario 2 ($Z = -6.7$, $p < .001$). The kappa coefficients for medians were not different between groups for scenarios 3 ($Z = -0.87$, $p = 0.386$), 4 ($Z = 0.06$, $p = .95$) and 6 ($Z = 1.22$, $p = .224$). The kappa coefficients for Structured GAS range scores were higher than those for Flexible GAS for scenarios 1 ($Z = 7.34$, $p < .001$) and 5 ($Z = 6.45$, $p < .001$). The kappa coefficients for ranges were not different between groups for scenarios 2 ($Z = -1.51$, $p = .13$), 3 ($Z = -0.118$, $p = .906$), 4 ($Z = 1.63$, $p = .103$) and 6 ($Z = 1.121$, $p = .262$).

DISCUSSION

The present study aimed to investigate and compare the inter-user reliability and temporal stability of conventional (Flexible) and modified (Structured) GAS scale construction in novice users.

Inter-user Reliability

As hypothesized, Structured GAS exhibited higher inter-user reliability than Flexible GAS for both the median and range values. With higher agreement between median scores ($\kappa = 0.28$), more participants in the Structured GAS group entered the same expected (0 level GAS) goal attainment value across scenarios than participants in Flexible GAS ($\kappa = 0.19$). This was also true for the range, with Structured GAS participants constructing GAS scales with more consistent ranges ($\kappa = 0.30$) than Flexible GAS ($\kappa = 0.05$). This demonstrated that among novice users of GAS, Structured GAS elicited more numerically similar GAS scales than Flexible GAS. Whilst increased experience with GAS is likely to improve inter-user reliability of scale construction, as has been put forward by Cytrynbaum et al. (1979) and Steenbeek et al. (2010) for inter-rater reliability, we have demonstrated that a more structured approach to constructing GAS scales is associated with superior inter-user reliability. This finding can largely be attributed to the use of the calculator and the accompanying instructions that required participants to simply enter two values (CLF and MRLF) into the calculator to create the GAS scales. Rehabilitation clinicians and clients alike may benefit from the use of a simpler process of constructing GAS scales than the conventional approach (Steenbeek et al., 2007; Bouwens et al., 2009).

Temporal Stability

In partial support of the second hypothesis, Structured GAS demonstrated higher temporal stability for more of the goal setting scenarios than Flexible GAS for both the median and range GAS values. Furthermore, five of six scenarios for Structured GAS yielded statistically significant kappa values for median scores compared with four of six scenarios for Flexible GAS; and five of six scenarios for Structured GAS yielded statistically significant kappa values for range values compared with one of six scenarios for Flexible GAS. For the median score, Structured GAS scales were more stable than Flexible GAS scales in two of the scenarios (1 and 5) and less stable in one (2). For the range values, Structured GAS scales were more stable than Flexible GAS scales in two of the scenarios (1 and 5). Curiously, there were three instances of the Kappa paradox found in scenario 1 for Flexible GAS (for the medians analysis) and scenario 2 for Structured GAS (for the medians and ranges analyses), where an extremely low Kappa value for the median score was calculated despite high observed and expected agreement between participants. This situation is explained by Sim and Wright (2005, p. 261), who state that if the prevalence of similar ratings is high, "the chance agreement is also high and kappa is reduced accordingly". In the context of this study, if the prevalence of identical participants' medians and ranges are high, this can artificially and paradoxically reduce the Kappa value. This was also supported by an analysis by Falotico and Quatto (2015), who retrieved a Kappa value of -0.2 when five out of six raters demonstrated perfect agreement. The authors go on to propose a permutation technique that yields a "robust Kappa" value which is purportedly more resistant to prevalence bias than the original Kappa value (Falotico & Quatto, 2015, p.467). Future studies may benefit from implementing this technique, although it is "computationally expensive" (Falotico & Quatto, 2015, p. 468). Our study was the first to investigate temporal stability of GAS scale construction by asking participants to set GAS scales based on controlled vignettes on two occasions. This methodology was feasible in the current study and could be applied in future studies examining temporal stability of GAS scale construction.

Limitations

The current study had a number of limitations. Firstly, there were substantially fewer participants at Time 2 compared to Time 1, as it was not compulsory to complete the second session and there was at least a two-week waiting period between sessions. This likely resulted in reduced power to show an effect of temporal stability. A considerable proportion of participants erroneously completed certain scenarios, likely due to their nature and wording. For example, a scenario about waking up early was interpreted in two ways, with participants either setting goals that required the subject to wake up at an earlier time, e.g. '7 am', or by frequency, 'wake up early five days a week'. This presented a challenge in standardizing the values for analysis. Responses that dealt with time instead of frequency were given arbitrary values (e.g. '57') that would be matched for other participants who provided the same answer. This was appropriate as the Kappa test was only concerned with the absolute agreement of values to determine agreement. Future studies examining GAS inter-user reliability and temporal stability of scale construction should ensure that the wording of the goal setting scenarios do not elicit responses on different dimensions. Another limitation of the current study is that the wording of the goals was not compared as in previous studies (May-Benson et al., 2021; Rushton et al., 2002). Future studies investigating the reliability of GAS scale construction should consider both the wording and numerical data used to construct the scales.

The interpretation of the Kappa statistic is a contentious topic in the literature. The suggested benchmarks used in this study were arbitrarily set by the authors Landis and Koch (1977), and other methods of interpreting Kappa values exist (Sim & Wright, 2005; Powers, 2012). Also, the magnitude of the Kappa statistic is affected by the number of categories (the levels of the GAS scale) and participants, with fewer categories yielding higher Kappa scores (Sim & Wright, 2005). The paradoxes seen in Scenario 2 for temporal stability are due to the prevalence bias that affect both Cohen's and Fleiss' Kappa (Sim & Wright, 2005; Falotico & Quatto, 2015). Future studies should test the permutation method proposed by Falotico and Quatto (2015) to rectify Kappa paradoxical behaviour.

Implications

GAS scales constructed using the modified Structured GAS approach, with higher inter-user reliability and temporal stability, are likely to result in higher chances of meeting the following Krasny-Pacini et al. (2016) criteria: equidistance of levels, attainability/difficulty, precise description of all levels, and unidimensionality. The Structured GAS calculator ensures that all GAS levels are approximately equidistant. The fact that the calculator sets the expected (GAS = 0) outcome at approximately the mid-way point between current functioning and maximum realistic functioning makes it more likely that the expected outcome will be neither too easy nor difficult to attain. From a quantitative perspective, Structured GAS ensures precision of quantification of each GAS level. The very fact that Structured GAS requires the goal setter to quantify a single variable ensures unidimensionality. In addition to meeting these criteria, Structured GAS also addresses the following checklist items from Grant and Ponsford (2014): all five levels are mutually exclusive, the five outcome levels are continuous, and all possible outcomes have been considered. The relatively weak agreement in GAS ranges across raters and time for Flexible GAS implies that the range of all possible outcomes using that approach is more variable and idiosyncratic.

GAS scales with demonstrably higher inter-user reliability and temporal stability may be used to compare GAS

outcomes across studies, particularly when investigating the same or similar populations. The modified GAS approach described in Berry et al. (2023) makes use of goal menus, which is a further aspect of standardization that might impart greater consistency in how GAS scales are constructed, which would also potentially allow comparison of goal outcomes across studies. The need for greater standardization to address threats to the validity and reliability of GAS has been raised in the literature (Logan et al., 2022).

The Krasny-Pacini et al (2016) guidelines do not include criteria for either of the clinimetric properties investigated in the current study, but may benefit from the same. As they have asserted, GAS clinimetric qualities depend mainly on how experienced the team is in GAS writing. If GAS scale construction can be made more consistent across users and time, the need for independent raters to ensure GAS scales constructed by one team are comparable to those made by others is obviated. If naive first year university students were able to formulate more reliable Structured than Flexible GAS scales, then it is reasonable to assume that clinicians and therapists who are either familiar or unfamiliar with GAS would be able to do the same for their patients. With greater inter-user reliability and temporal stability, Structured GAS appears to demonstrate preliminary feasibility for use in clinical populations (Steenbeek et al., 2007; Ruble et al., 2012). The current study expands on a demonstration of preliminary feasibility, reliability and validity of Structured GAS in a substance use disorder rehabilitation setting (Berry et al., 2023). Future studies should investigate these and other clinimetric properties of Structured GAS in varied clinical / rehabilitation contexts.

CONCLUSION

The validity of GAS in clinical settings has been debated in the literature due to its variable methodology (Krasny-Pacini et al., 2016). The purpose of this study was to investigate and compare the inter-user reliability and temporal stability of two approaches to GAS scale construction. The results of this study indicated that the novel, calculator-based modified version of GAS demonstrated greater inter-user reliability and temporal stability of scale construction than conventional GAS. Although these results were based on the responses to a predetermined set of goal-setting scenarios among novice users in a non-rehabilitation context, they support the utility of investigation of these clinimetric properties of modified GAS in clinical contexts.

NOTE

The data that support the findings of this study are available from the corresponding author upon reasonable request.

FUNDING

This research did not receive any specific funding.

ACKNOWLEDGEMENTS

The authors would like to thank the participating students for their time and engagement in this study.

REFERENCES

1. Ashford, S., Jackson, D., & Turner-Stokes, L. (2015). Goal setting, using goal attainment scaling, as a method to identify patient selected items for measuring arm function. Physiotherapy, 101(1), 88-94. https://doi:10.1016/j.physio.2014.04.001
2. Berry, J., Marceau, E. M., & Lunn, J. (2023). Feasibility, reliability and validity of a modified approach to goal attainment scaling to measure goal outcomes following cognitive remediation in a residential substance use disorder rehabilitation setting. Australian Journal of Psychology, 75(1). https://doi-org.simsrad.net.ocs.mq.edu.au/10.1080/00049530.2023.2170652
3. Bouwens, S. F., Van Heugten, C. M., & Verhey, F. R. (2009). The practical use of goal attainment scaling for people with acquired brain injury who receive cognitive rehabilitation. Clinical Rehabilitation, 23(4), 310-320. https://doi: 10.1177/0269215508101744
4. Bovend'Eerdt, T. J., Botell, R. E., & Wade, D. T. (2009). Writing SMART rehabilitation goals and achieving goal attainment scaling: a practical guide. Clinical rehabilitation, 23(4), 352-361. https://doi:10.1177/0269215508101741
5. Bovend'Eerdt, T. J., Dawes, H., Izadi, H., & Wade, D. T. (2011). Agreement between two different scoring procedures for goal attainment scaling is low. Journal of Rehabilitation medicine, 43(1), 46-49. https://doi:10.2340/16501977-0624
6. Cheema, K., Dunn, T., Chapman, C., Rockwood, K., Howlett, S. E., & Sevinc, G. (2024). A systematic review of goal attainment scaling implementation practices by caregivers in randomized controlled trials. Journal of Patient-Reported Outcomes, 8(1), 37. https://doi: 10.1186/s41687-024-00716-w. PMID: 38530578; PMCID: PMC10965877
7. Clark, M., Miller, A., Berry, J., & Cheng, K. (2021). Mental contrasting with implementation intentions increases study time for university students. The British journal of educational psychology, 91(3), 850–864. https://doi.org/10.1111/bjep.12396
8. Cornbleth, T. (1978) Evaluation of Goal Attainment in Geriatric Settings. Journal of the American Geriatrics Society, 26: 404-407. https://doi-org.simsrad.net.ocs.mq.edu.au/10.1111/j.1532-5415.1978.tb05386.x

9. Cox, R., & Amsters, D. (2002). Goal attainment scaling: an effective outcome measure for rural and remote health services. Australian Journal of Rural Health, 10(5), 256- 261. https://doi:10.1111/j.1440-1584.2002.tb00041.x
10. Cusick, A., McIntyre, S., Novak, I., Lannin, N., & Lowe, K. (2006). A comparison of goal attainment scaling and the Canadian Occupational Performance Measure for paediatric rehabilitation research. Pediatric Rehabilitation, 9(2), 149-157. https://doi:10.1080/13638490500235581
11. Cytrynbaum, S., Ginath, Y., Birdwell, J., & Brandt, L. (1979). Goal attainment scaling: A critical review. Evaluation Quarterly, 3(1), 5-40. https://doi:10.1177%2F0193841X7900300102
12. Engelen, V., Ketelaar, M., & Gorter, J. W. (2007). Selecting the appropriate outcome in paediatric physical therapy: how individual treatment goals for children with cerebral palsy are reflected in GMFM-88 and PEDI. Journal of Rehabilitation Medicine, 39(3), 225-231. https://doi:10.2340/16501977-0040
13. Evans, J. J. (2012). Goal setting during rehabilitation early and late after acquired brain injury.
14. Current Opinion in Neurology, 25(6), 651-655. https://doi:10.1097/WCO.0b013e3283598f75
15. Falotico, R., & Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. Quality & Quantity, 49(2), 463-470. https://doi:10.1007/s11135-014-0003-1
16. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378. https://doi:10.1.1.456.3830
17. Fleiss, J. L., Levin, B., & Paik, M. C. (1981). The measurement of interrater agreement.
18. Statistical Methods for Rates and Proportions, 2(212-236), 22-23. https://doi:10.1.1.456.3830
19. Grant, M., & Ponsford, J. (2014). Goal attainment scaling in brain injury rehabilitation: Strengths, limitations and recommendations for future applications. Neuropsychological Rehabilitation, 24(5), 661-677. https://doi:10.1080/09602011.2014.901228
20. Heavlin, W. D., Lee-Merrow, S. W., & Lewis, V. M. (1982). The psychometric foundations of goal attainment scaling. Community Mental Health Journal, 18(3), 230-241. https://doi:10.1007/BF00754339
21. Hurn, J., Kneebone, I., & Cropley, M. (2006). Goal setting as an outcome measure: a systematic review. Clinical Rehabilitation, 20(9), 756-772. https://doi:10.1177/0269215506070793
22. Khan, F., Pallant, J. F., & Turner-Stokes, L. (2008). Use of goal attainment scaling in inpatient rehabilitation for persons with multiple sclerosis. Archives of Physical Medicine and Rehabilitation, 89(4), 652-659. https://doi:10.1016/j.apmr.2007.09.049
23. King, G. A., McDougall, J., Palisano, R. J., Gritzan, J., & Tucker, M. A. (2000). Goal attainment scaling: its use in evaluating pediatric therapy programs. Physical & Occupational Therapy in Pediatrics, 19(2), 31-52. https://doi:10.1080/J006v19n02_03
24. Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. Community Mental Health Journal, 4(6), 443-453. https://doi:10.1007/BF01530764
25. Kiresuk, T. J. (1973). Goal attainment scaling at a county mental health service. Evaluation: The International Journal of Theory, Research and Practice, 12-18.
26. Koski, J., & Richards, L. G. (2015). Reliability and Sensitivity to Change of Goal Attainment Scaling in Occupational Therapy Nonclassroom Educational Experiences. The American journal of occupational therapy : official publication of the American Occupational Therapy Association, 69 Suppl 2, 6912350030p1–6912350030p5. https://doi.org/10.5014/ajot.2015.016535
27. Krasny-Pacini, A., Hiebel, J., Pauly, F., Godon, S., & Chevignard, M. (2013). Goal attainment scaling in rehabilitation: a literature-based update. Annals of Physical and Rehabilitation Medicine, 56(3), 212-230. https://doi:10.1016/j.rehab.2013.02.002
28. Krasny-Pacini, A., Evans, J., Sohlberg, M. M., & Chevignard, M. (2016). Proposed criteria for appraising goal attainment scales used as outcome measures in rehabilitation research. Archives of Physical Medicine and Rehabilitation, 97(1), 157-170. https://doi:10.1016/j.apmr.2015.08.424
29. Laerd Statistics (2019). Fleiss' Kappa using SPSS Statistics. Statistical Tutorials and Software Guides. Retrieved August 14, 2020, from https://statistics.laerd.com/spss-tuorials/fleiss- kappa-in-spss-statistics.php
30. Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data.
31. Biometrics, 33(1), 159-174. https://doi:10.2307/2529310
32. Logan, B., Jegatheesan, D., Viecelli, A., Pascoe, E., & Hubbard, R. (2022). Goal attainment scaling as an outcome measure for randomised controlled trials: A scoping review. BMJ Open, 12(7), e063061. https://doi.org/10.1136/bmjopen-2022-063061
33. MacKay, G., Somerville, W., & Lundie, J. (1996). Reflections on goal attainment scaling (GAS): Cautionary notes and proposals for development. Educational Research, 38(2), 161-172.
34. https://doi:10.1080/0013188960380204
35. May-Benson, T. A., Schoen, S. A., Teasdale, A., & Koomar, J. (2021). Inter-Rater Reliability of Goal Attainment Scaling with Children with Sensory Processing Disorder. The Open Journal of Occupational Therapy, 9(1), 1–13. https://doi.org/10.15453/2168-6408.1693
36. McLaren, C., & Rodger, S. (2003). Goal attainment scaling: Clinical implications for paediatric occupational therapy practice. Australian Occupational Therapy Journal, 50(4), 216-224. https://doi.org/10.1046/j.1440-1630.2003.00379.x
37. Ottenbacher, K. J., & Cusick, A. (1993). Discriminative versus evaluative assessment: Some observations on

38. goal attainment scaling. American Journal of Occupational Therapy, 47(4), 349-354.
39. https://doi:10.5014/ajot.47.4.349
40. Palisano, R. J. (1993). Validity of goal attainment scaling in infants with motor delays. Physical Therapy, 73(10), 651-658. https://doi.org/10.1093/ptj/73.10.651
41. Powers, D. M. W. (2012). The problem with kappa. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 345-355. https://doi:10.5555/2380816.2380859
42. Rockwood, K., Joyce, B., & Stolee, P. (1997). Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. Journal of Clinical Epidemiology, 50(5), 581-588. https://doi:10.1016/S0895-4356(97)00014-0
43. Rushton, P. W., & Miller, W. C. (2002). Goal attainment scaling in the rehabilitation of patients with lower-extremity amputations: A pilot study. Archives of Physical Medicine and Rehabilitation, 83(6), 771–775. https://doi.org/10.1053/apmr.2002.32636
44. Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. Psychological Bulletin, 111(2), 352. https://doi:10.1037/0033-2909.111.2.352
45. Schlosser, R. W. (2004). Goal attainment scaling as a clinical measurement technique in communication disorders: a critical review. Journal of communication disorders, 37(3), 217-239. https://doi:10.1016/j.jcomdis.2003.09.003
46. Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy, 85(3), 257-268. https://doi:10.1093/ptj/85.3.257
47. Shankar, S., Marshall, S. K., & Zumbo, B. D. (2020). A Systematic Review of Validation Practices for the Goal Attainment Scaling Measure. Journal of Psychoeducational Assessment, 38(2), 236-255. https://doi-org.simsrad.net.ocs.mq.edu.au/10.1177/0734282919840948
48. Steenbeek, D., Meester-Delver, A., Becher, J. G., & Lankhorst, G. J. (2005). 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. Clinical Rehabilitation, 19(3), 274-282. https://doi:10.1191%2F0269215505cr859oa
49. Steenbeek, D., Ketelaar, M., Galama, K., & Gorter, J. W. (2007). Goal attainment scaling in paediatric rehabilitation: a critical review of the literature. Developmental Medicine & Child Neurology, 49(7), 550-556. https://doi:10.1111/j.1469-8749.2007.00550.x
50. Steenbeek, D., Ketelaar, M., Lindeman, E., Galama, K., & Gorter, J. W. (2010). Interrater reliability of goal attainment scaling in rehabilitation of children with cerebral palsy. Archives of Physical Medicine and Rehabilitation, 91(3), 429-435. https://doi:10.1016/j.apmr.2009.10.013
51. Steenbeek, D., Gorter, J. W., Ketelaar, M., Galama, K., & Lindeman, E. (2011). Responsiveness of Goal Attainment Scaling in comparison to two standardized measures in outcome evaluation of children with cerebral palsy. Clinical rehabilitation, 25(12), 1128-1139. https://doi:10.1177/0269215511407220
52. Stevens, S. (1946). On the Theory of Scales of Measurement. Science, 103(2684), 677-680.
53. Stolee, P., Rockwood, K., Fox, R. A., & Streiner, D. L. (1992). The use of goal attainment scaling in a geriatric care setting. Journal of the American Geriatrics Society, 40(6), 574-
54. 578. https://doi:10.1111/j.1532-5415.1992.tb02105.x
55. Tennant, A. (2007). Goal attainment scaling: current methodological challenges. Disability and Rehabilitation, 29(20-21), 1583-1588. https://doi:10.1080/09638280701618828
56. Turner-Stokes, L. (2009). Goal attainment scaling (GAS) in rehabilitation: a practical guide. Clinical Rehabilitation, 23(4), 362-370. https://doi:10.1177%2F0269215508101742
57. Vu, M., & Law, A. V. (2012). Goal-attainment scaling: a review and applications to pharmacy practice. Research in Social and Administrative Pharmacy, 8(2), 102-121. https://doi:10.1016/j.sapharm.2011.01.003
58. Yip, A. M., Gorman, M. C., Stadnyk, K., Mills, W. G., MacPherson, M., & Rockwood, K. (1998). A standardized menu for Goal Attainment Scaling in the care of frail elders. The Gerontologist, 38(6), 735-742. https://doi:10.1093/geront/38.6.735