J Autism Dev Disord (2012) 42:1974–1983
DOI 10.1007/s10803-012-1446-7

ORIGINAL PAPER

Goal Attainment Scaling as an Outcome Measure in Randomized Controlled Trials of Psychosocial Interventions in Autism

Lisa Ruble · John H. McGrew · Michael D. Toland

Published online: 21 January 2012

Abstract Goal attainment scaling (GAS) holds promise as an idiographic approach for measuring outcomes of psychosocial interventions in community settings. GAS has been criticized for untested assumptions of scaling level (i.e., interval or ordinal), inter-individual equivalence and comparability, and reliability of coding across different behavioral observation methods. We tested assumptions of equality between GAS descriptions for outcome measurement in a randomized trial (i.e., measurability, equidistance, level of difficulty, comparability of behavior samples collected from teachers vs. researchers and live vs. videotape). Results suggest GAS descriptions can be evaluated for equivalency, that teacher collected behavior samples are representative, and that varied sources of behavior samples can be reliably coded. GAS is a promising measurement approach. Recommendations are provided to ensure methodological quality.

Keywords Goal attainment scaling · Outcome measurement · Autism · Randomized controlled trials · Reliability · Psychosocial intervention

Although past and current legislation on educational reform (No Child Left Behind Act 2001; Individuals with Disabilities Act 2004) continues to focus on outcome measurement, the academic and standards-based assessment and accountability measurement systems used in today's classrooms are limited to monitoring of academic skills when applied at the program or group levels of assessment (i.e., nomothetic assessment). However, these systems fall short of what is needed in an assessment system at the level of the individual student—the ability to discern whether or not children with individualized teaching plans are responding to their educational programs (i.e., idiographic assessment). For more than 30 years, educational researchers and practitioners have lamented the limitations of traditional assessment methods for monitoring the quality and impact of educational programs of children with disabilities (Shuster et al. 1984). Similar challenges have been identified by researchers and clinicians from other fields when evaluating the quality and outcomes of their services for persons receiving individualized programming. Most of our current assessment systems, including curriculum based measurement approaches, continue to assume a nomothetic approach, that all individuals can be assessed along similar dimensions, using similar measures or rating systems. Nomothetic approaches work well for group comparisons, when everyone can be viewed as or treated as having similar goals (as in a class of students all learning the same curriculum), but lack sensitivity for classroom and large scale evaluation approaches to assist teachers and administrators in making data based decisions regarding the effectiveness of their instruction at the student level (Quenemoen et al. 2003) and for those in special education programs. Within special education, by definition, each student has an individualized plan, and, by law, the assessment of that plan must take into account the fact that goals and how to rate them are also individualized. This is the definition of an idiographic approach. Progress monitoring systems, such as general curriculum based measurement, also pose limitations due to a lack of standards for the

L. Ruble (✉) · M. D. Toland
Department of Educational, School, and Counseling Psychology, University of Kentucky, 237 Dickey Hall, Lexington, KY 40506-0017, USA
e-mail: lisa.ruble@uky.edu

J. H. McGrew
Department of Psychology, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

nonacademic skills that often are central to the special education curriculum. For example, the educational programs of students with autism must include specialized individualized instruction on communication, socialization, and independence (NRC 2001)—pivotal skills that underlie success in all areas of learning and are associated with positive outcomes, yet difficult to measure. Alternative measurement approaches are, therefore, necessary and crucial for monitoring progress and measuring outcomes of essential skills for students in special education, such as those with autism.

Goal attainment scaling emerged more than 40 years ago as one possible solution to the need for individualized assessment. Originally developed as a measurement tool first applied in mental health settings (Cytrynbaum et al. 1979; Kiresuk and Sherman 1968; Kiresuk et al. 1994), goal attainment scaling was created for program evaluation purposes and was considered superior for its suitability for individual and group outcome analysis of highly diverse and individualized treatments (Kiresuk and Sherman 1968). Since then goal attainment scaling has become a standard outcome measurement approach for school consultation research (e.g., Ruble et al. 2010a; Sheridan et al. 2006; Sladeczek et al. 2001), particularly because it is compatible with the IEP objectives that operationalize a student's special education program goals (Oren and Ogletree 2000; Shuster et al. 1984).

Although goal attainment scaling holds promise as a substantive and sensitive approach for measuring the outcomes of personalized psychosocial interventions delivered in mental health and educational settings, there are several assumptions of this approach that have not been fully explored empirically, which have served to limit its acceptability and applicability in research and evaluation. Specific concerns include whether GAS scores are interval or ordinal, are comparable across groups, and are reliable and comparable when people apply different behavioral observation methods (e.g., ratings from live observation vs. videotaped observation or from teacher supplied vs. researcher supplied tapes of behavior).

The first issue, level of measurement, has sparked considerable debate about the computation and statistical analyses of GAS scores. MacKay et al. (1996) argued strongly that GAS scores do not represent interval data and should not be converted into standard scores, encouraging instead the use of non-parametric approaches. In contrast, others have argued that parametric methods are appropriate because the metric approximates a normal distribution, and results based on assuming the data are ordinal or interval show negligible differences (Cardillo and Smith 1994; Malec 1999; Ottenbacher and Cusick 1993). GAS scores are most likely to approximate interval ratings when benchmarks are carefully constructed based on clear procedures and include consistent and well defined use of frequency or other qualitative features (e.g., prompting level). Moreover, it is sometimes difficult to insure that ratings are truly of equal intervals (i.e., some ratings may be rather close to one another or perhaps too far apart). At the very least, these problems and controversies suggest that researchers should exercise care in the creation of GAS and apply the simplest and most practically meaningful approach of using raw scores (i.e., the total score across goals or the average rating across goals) for comparisons between groups rather than converted scores based on assumptions of normality. This is because using raw total scores across goals creates more score categories which tend to more likely reflect a continuous distribution and better reflect a normal distribution.

The second and third issues of comparability of GAS scores across groups and interrater agreement when behavior is collected using different sources were particularly important for our purposes. Goal attainment scaling was applied as the outcome measurement approach in two randomized controlled trials (RCT; Ruble et al. 2010a, b, 2011) to evaluate a parent-teacher consultation program and decision-planning framework called the Collaborative Model for Promoting Competence and Success (COMPASS; Ruble and Dalrymple 2002; Ruble et al. 2012) specialized to children with autism. COMPASS helps identify educational goals and objectives and generates personalized, research supported, educational intervention strategies using information provided by parents and teachers on the personal and environmental challenges and supports of the child with autism. The COMPASS intervention consisted of an initial 3-h parent-teacher consultation that took place near the start of the school year, followed by four teacher coaching sessions that lasted about 90 min each and occurred about 4–6 weeks throughout the remainder of the school year. Goal attainment scaling was the primary monitoring and outcome measurement tool applied throughout all study phases of baseline, coaching, and final evaluation. Both studies included a comparison or placebo control group. The second study included a third condition consisting of a web-based teacher coaching group, which utilized web-based videoconferencing technology for the follow-up coaching sessions to compare the effects of web-based versus face-to-face coaching (Ruble et al. 2011). For both studies, the final GAS scores were collected at the end of the school year by an observer who used direct observation rather than teacher report, was independent from the research team, and was unaware of group assignment.

A serious concern when comparing groups in randomized trials is the need to demonstrate the pre-intervention equivalence of GAS descriptions between the control and intervention groups. If GAS scores are higher in the experimental conditions at the end of the school year,

an alternative explanation might be that the initial scaled descriptions of behavior were not equivalent across groups. That is, one could argue that the targeted outcomes as scaled using GAS were less difficult and easier for children in the experimental group to achieve compared to the control group children; 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. Although ideally comparability should be realized when utilizing random assignment, it should not be assumed in experimental or other correlational research designs.

To address these potential issues of comparability, we developed a coding framework for GAS scoring and operationalized three comparison features: (a) measurability; (b) equidistance; and (c) level of difficulty. We asked three questions: (a) is the goal and the associated benchmarks relative to each goal described in measurable terms that are comparable between groups; (b) is the distance between each of the benchmarks for each scale of equal intervals and comparable between groups; and (c) is the level of difficulty between the baseline or starting levels of performance and the targeted outcome goal comparable between groups? If differences are identified, then outcome analyses should account for them statistically using the scores as covariates. We were also interested in the independence of these three comparison features and asked if they correlated with one another.

For the third issue of reliability, we were interested in whether interrater agreement was impacted by method of observation (i.e., use of live vs. videotaped behavior samples). If we could demonstrate adequate interrater agreement from the same behavior coded from direct observation and from videotape, then we would have confidence in the use of videotaped recordings for monitoring GAS scores. Also, the ability to use teacher videotaping would reduce the likelihood of missing data and free up a practically significant amount of consulting time as teachers could submit videotapes of instruction that would allow for valid and reliable data and serve as an acceptable substitute for a consultant or researcher going into the classroom for data collection purposes. It would also keep children from needing to be disrupted by intrusive observation from their classroom routine, reducing stress for the teacher and student. We were also interested in whether scores would be equivalent if behavioral samples were selected by different sources (i.e., teacher- or researcher-selected behavior samples). For example, were tapes provided by teachers biased toward showing the best case scenarios of behavior, potentially causing score inflation? Because the second study focused on teacher provided examples of behavior during the four follow-up teacher coaching sessions, a potential concern was that teachers taped instructional situations repeatedly until they collected the most optimum example that was not truly representative of the child's actual and most consistent level of performance.

To address the aforementioned issues, we conducted two sets of analyses. To test for observer differences in choosing behavior samples, using study 2 data, we compared the mean GAS scores collected from behavior samples provided by the teacher to samples collected by direct observations of instruction conducted on a different day by the researcher. Both teacher and researcher behavior samples were based on the same targeted skills. To test for differences due to videotaping, also in study 2, we evaluated the interrater agreement between a coder who rated behavior from direct observation and a second coder who rated the same behavior sample from a videotape collected by the researcher conducting the live observation.

In summary, for the first set of questions, we asked: (a) are the three features—measurability, equidistance, and level of difficulty of the targeted outcome objectives similar between the comparison and experimental groups, and (b) independent from one another? For the second set of questions, we asked: (a) are scores based on teacher supplied versus researcher supplied videotaped samples of behavior similar (i.e., same goal, but two different behavior samples collected on different days but within 1 week of each other), and (b) are GAS ratings coded from live versus video-taped samples of the same behavior (measured at the same time) similar?

Method

Both the first and second set of questions were answered using secondary analysis of data collected from two randomized controlled trials (RCTs) of teacher consultation for students with autism (Ruble et al. 2010a, 2011).

Overview of Study Sample 1 and Study Sample 2

Participants and Procedure

For both study samples, similar procedures were used for participant recruitment and selection. Teacher participants were special education teachers who had at least one student with autism on her caseload. Child participants were selected if they were between the ages of 3–8 years, had an existing DSM-IV-TR diagnosis of autism and no comorbid sensory disorders, and were receiving special education services in a public school under the educational category of autism. The researchers contacted district and school administrators in one Southeastern and one Midwestern

state for permission and support to recruit teachers. Interested administrators forwarded names of special education teachers to the researchers. Then, special education teachers were contacted directly by the researchers. Following teacher agreement to participate, the initials of all students with autism were provided, and one child was randomly selected from the teacher's caseload. The teacher then asked the parent/caregiver of that student for permission to be contacted by the researchers. If the parent/caregiver refused to participate, another child was randomly selected. Both teachers and parents/caregivers provided informed consent to participate.

Child participants received one of two screeners—the Modified Checklist for Autism in Toddlers (M-CHAT; Robins et al. 2001) for children under 4 years old, and the Social Communication Questionnaire (SCQ; Rutter et al. 2004) for those 4 or older. The diagnosis of autism was confirmed with the Autism Diagnostic Observation Schedule—Generic (Lord et al. 2000). In addition, standardized measures of child language (Oral and Written Language Scales; Carrow-Woolfolk 1995), cognitive ability (Differential Abilities Scale; Elliott 1990), and adaptive behavior (Classroom Edition of the Vineland Adaptive Behavior Scales; Sparrow et al. 2005) were also used to determine group equivalence between participants prior to randomization and at baseline.

Participants recruited for the first study were 35 teacher-student dyads and for the second study 44 teacher-child dyads. After random assignment, for the first study 17 dyads completed the control condition and 18 completed the face-to-face condition; for the second study, 15 dyads completed a placebo control condition, in which teachers participated in online instructional modules on three evidence based practices in autism; 14 completed face-to-face coaching sessions; and 15 completed web-based coaching sessions. For the two experimental conditions, both received an initial face-to-face COMPASS consultation near the beginning of the school year. The demographic characteristics of participants for each study are provided in Table 1.

Table 1 Participant characteristics of sample 1 and 2

	Sample 1 *M* (*SD*)	Sample 2 *M* (*SD*)
Child		
Age (years)	6.1 (1.7)	5.7 (1.5)
Differential Abilities Scale[a]	46.8 (24.1)	56.3 (22.1)
Oral and Written Language Scales[a]	46.7 (18.5)	53.6 (13.1)
Vineland Adaptive Behavior Scales (TR)[b]	63.6 (13.3)	59.8 (13.7)
Teacher		
Total number of children with autism taught[c]	6.5 (9.0)	8.4 (16.4)
Total years with children with autism	6.8 (7.0)	5.7 (5.7)

TR teacher report

[a] Standard score

[b] Teacher report

[c] Refers to total throughout teaching career

Goal Attainment Scale Development and Measurement

Studies on interrater agreement show that GAS ratings are reliable when objectives are clear and measurable (Cytrynbaum et al. 1979; Malec 1999; Schlosser 2004; Shefler et al. 2001; Stolee et al. 1999). To ensure generation of clear and measureable goals, a written procedure for creating the GAS descriptions were used for both study 1 and 2 (see Table 2, Fig. 1; Ruble et al. 2012). After the GAS descriptions were written, an independent observer coded each goal for the three features of measurability, equidistance, and difficulty. Table 2 and Fig. 1 summarize instructions used by observers for coding GAS descriptions. As mentioned earlier, because study 2 included web-based videoconferencing, we also examined potential differences in interrater agreement of behavior samples coded live versus via videotape, and comparability when ratings were made from tapes collected and supplied by teachers versus those obtained by the research team.

Goal attainment scale measurement of student progress of the experimental groups was evaluated against student progress of the control groups toward three IEP objectives selected at the start of the school year. IEP objectives were selected if they represented skills critical for students with autism: a social skill, a communication skill, and a learning skill for both groups (NRC 2001; Ruble et al. 2010b). For the first study sample, the control group represented children who received their usual special education program. Control group teachers in the second study sample were provided online resources of evidence based teaching methods for students with autism (e.g., http://autismpdc.fpg.unc.edu/content/evidence-based-practices) that are available to the public. Both were considered control group participants because they had no direct interaction with the research team throughout the school year.

The experimental groups, however, all received the consultation intervention, COMPASS. Details of the intervention are provided in Ruble et al. (2010a, 2012). An initial consultation was provided to the child's teacher and parent at the beginning of the school year. One outcome of the consultation was the selection of three personalized teaching goals and intervention plans for each child. The domains of the teaching goals were similar to the control group and represented a social, communication, or independent skill.

Table 2 Definitions of measurability, difficulty, and equidistance ratings used by independent raters

Measurability
1. None or only one indicator (prompt level, criterion for success; observable skill) is described in the goals
2. Two of the three indicators (prompt level, criterion for success; observable skill) are described in the goals
3. All three indicators (prompt level, criterion for success; observable skill) are described in the goals
[Note: For prompt level, both the type of prompt and frequency of prompt must be provided in all the goals for the highest score]
Difficulty
1. Skill is very close to what the child is already described as able to perform in the present levels of performance
2. The present levels of performance indicates that the child is able to perform the skill in limited ways compared to what is written in the objective (limited people, prompts, or places); if present levels says the child has difficulty with the skill, score a "2"
3. The present levels of performance indicates that the child is unable to perform skill with anyone, anywhere, or with any prompts compared to what is written in the objective
Equidistance
1. None or only one of the three descriptions are equilibrated appropriately in reference to the goal.
2. Two of the three descriptions are equilibrated appropriately in reference to the goal.
3. All of the three descriptions relative to the goal are equilibrated and scaled appropriately
[Note: Refer to Fig. 1 for examples of ordered criteria. Prompts are correctly ordered when they go from most to least restrictive and/or the skill frequency increased by 50% relative to objective for +1 and +2 and is decreased by 50% for −1 (do not include the present levels of performance −2 description in the rating of this dimension)]

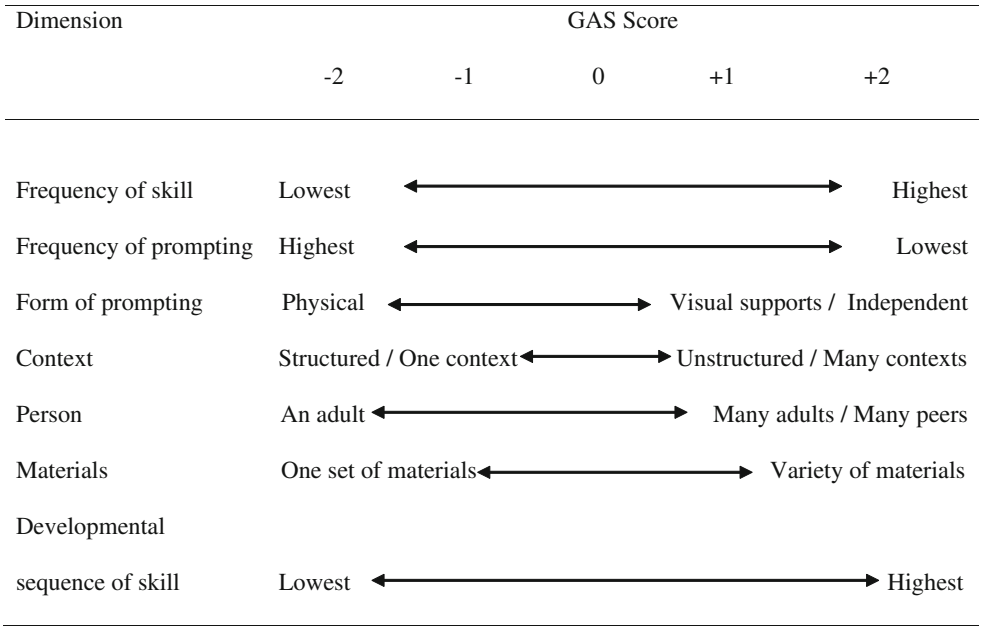

Fig. 1 Considerations when writing the GAS benchmarks

After the initial consultation, teachers received four coaching sessions throughout the remainder of the school year. Each coaching session occurred about every 5 weeks. The GAS templates were developed for the experimental groups subsequent to the initial consultation and prior to the first teacher coaching session.

As mentioned, to create the GAS descriptions for both experimental and control groups, a protocol was developed to ensure that the descriptions were written using a systematic approach that would facilitate comparability between groups (complete instructions are available from first author). The original process for developing and using goal attainment scaling is described in Kiresuk et al. (1994) and involves (a) creating the goal attainment scale (GAS) prior to the onset of intervention; (b) implementing the intervention; (c) evaluating progress following the intervention; and (d) comparing outcomes against targeted goals within and across individuals. Prior to the onset of intervention, goals are identified and written onto a GAS template (see Table 3). An individual may have a single goal or multiple goals. Weights can be assigned to each goal according to overall priority and relative to the other

Table 3 Example of a completed GAS form

−2 Present level of performance	−1 Progress	0 Expected level of outcome (GOAL)	+1 Somewhat more than expected	+2 Much more than expected
Aggresses when given a task he does not want to do. Is difficult to motivate. Does not have a more appropriate way to communicate refusals or to negotiate	When presented with a task menu, Anthony will start and complete three (1) 2–3 min tasks each day without aggression with one (2) adult verbal cue (e.g., time to work) and gestural/picture cues across 2 weeks	When presented with a task menu, Anthony will start and complete three 2–3 min tasks each day without aggression with one adult verbal cue (e.g., time to work) and gestural/picture cues across 2 weeks	When presented with a task menu, Anthony will start and complete three (4) 2–3 min tasks each day without aggression with one (0) adult verbal cue (e.g., time to work) and gestural/picture cues across 2 weeks	When presented with a task menu, Anthony will start and complete three (6) 2–3 min tasks each day without aggression with one (0) adult verbal cue (e.g., time to work) and gestural/picture cues across 2 weeks
Has difficulty imitating others, especially children using actions with objects. Likes objects he can manipulate	Anthony will imitate play activities for five (2) minutes with at least three (1) different preferred objects (dinosaurs, animals, doll…) each day across 2 weeks	Anthony will imitate adult play activities for 5 min with at least three different preferred objects (dinosaurs, animals, doll…) each day across 2 weeks	Anthony will imitate adult play activities for five (7) minutes with at least three (4) different preferred objects (dinosaurs, animals, doll…) each day across 2 weeks	Anthony will imitate adult (peer) play activities for five (10) minutes with at least three (6) different preferred objects (dinosaurs, animals, doll…) each day across 2 weeks
May use aggression as a way to request. Relies on adult prompts to make requests	Anthony will make 10 (5) different requests per day independently (with verbal cues) or as a response to a question (go home, eat, help, more, finished, various objects/activities) using sign, pictures, or verbal on a daily basis	Anthony will make 10 different requests per day independently (go home, eat, help, more, finished, various objects/activities) or as a response to a question ("what do you want?") using sign, pictures, or verbalization on a daily basis	Anthony will make 10 (15) different requests per day independently (go home, eat, help, more, finished, various objects/activities) or as a response to a question ("what do you want?") using sign, pictures, or verbalization on a daily basis	Anthony will make 10 (20) different requests per day independently (go home, eat, help, more, finished, various objects/activities) or as a response to a question ("what do you want?") using sign, pictures, or verbalization on a daily basis

objectives. To create the GAS, descriptors or benchmarks are developed for each of the criterion levels. The standard GAS measurement system is based on a 5-point response scale: −2 (worse expected outcome), −1 (less than expected outcome), 0 (expected outcome), +1 (more than expected outcome), and +2 (best expected outcome). After the GAS is generated, the intervention(s) is provided, and the GAS can be used to monitor progress toward goal completion at specified time durations or at the end of the intervention. For determining final progress toward goal attainment, raw scores based on the GAS are computed. Because scores are assumed to be standardized, they can be converted to T scores (i.e., $M = 50$; $SD = 10$) using the Kiresuk-Sherman formula (Kiresuk et al. 1994) and compared across individuals.

Instead of applying the standard 5-point description (i.e., −2 = worst possible outcome), we modified our scales to better represent our population and better match our evaluation approach. Because autism is not a disability associated with regression (unlike disorders such as muscular dystrophy or Rett's syndrome), we applied the recommendation from Schlosser (2004), that the −2 level represent baseline or present level of performance. If indeed, the child made no progress on their objective by the end of the school year, this would in fact represent the worst possible outcome. Thus, the following 5-point rating scale used was: −2 = child's present levels of performance, −1 = progress, 0 = expected level of outcome, +1 = somewhat more than expected, +2 = much more than expected (see Table 3). The score of zero represented improvement consistent with the actual description of the written IEP objective. The GAS scores for each of the three skills were summed at the end of the year following the COMPASS intervention. As mentioned, all of the GAS ratings used to determine post-treatment effectiveness was based on direct observations rather than teacher ratings. For final data assessment of study samples 1 and 2, teachers were instructed to demonstrate for the independent observer each of the three targeted objectives during a curriculum based instructional situation, which typically lasted for 20 min. To determine inter-rater agreement of the GAS scores, the mean sum of the total raw scores was calculated at baseline and at final evaluation using sample 1 data. The scores from the blinded observer were compared to the scores determined by a primary observer. Using intraclass correlation (ICC), the inter-rater agreement of the ratings for the

blinded final observer and the primary observer was .99 at the final assessment.

Data Analysis

We treated the mean of the sum of the raw GAS scores as interval variables because we developed and applied a standard approach for developing the GAS benchmarks that was designed to emulate equal interval scale construction, as recommended by Schlosser (2004). That is, the GAS descriptions were based on accumulative frequency, decreasing prompts, and increasing performance or generalization (i.e., able to demonstrate skill with more adults or peers, in different classrooms, and with different materials; see Fig. 1). For the first set of research questions Independent *t* tests were used to compare the mean scores on each of the three criterion features (measurability, equidistance, difficulty) by group assignment (control; experimental) using study sample 1, and ANOVAs were used to evaluate the comparability of the three features by group assignment (control; face-to-face; web-based) using study sample 2. Pearson correlations were used to test the relationship between each of the three criterion features within each sample.

For the second set of research questions we wanted to determine whether teachers provided a biased or inflated sample of behavior for coding. So, a paired *t* test was conducted to compare GAS ratings of behavior samples collected from videotaped samples from the teacher and researcher. A paired *t* test was also conducted to compare GAS ratings of behavior samples collected from live observations from the researcher and videotaped by the researcher. Intraclass correlations were also calculated to determine agreement and consistency between the different sources (i.e., teacher vs. researcher supplied videotape and live vs. videotaped).

Results

Interrater Reliability of Measurability, Equidistance and Level of Difficulty

After the GAS forms were written, a rater who was unaware of the group assignment evaluated the scaling for each goal on three criteria: (a) measurability of the scaled descriptions; (b) equidistance between each of the benchmark descriptions; and (c) degree of difficulty of the outcome skill compared to the student's present levels of performance. To establish interrater agreement, two raters independently coded 20% of the GAS forms for the three features of measurability, equidistance, and difficulty using sample 1. The intraclass correlation (ICC) for average agreement was .96 (95% CI [.87, .99]) for measurability, .96 (95% CI [.74, .99]) for equidistance, and .59 (95% CI [−.18, .81]) for difficulty. Using the same procedure for sample 2, the ICC for average agreement was 1.0 for measurability, .96 (95% CI [.84, .99]) for equidistance, and .96 (95% CI [.83, .99]) for difficulty.

Measurability, Equidistance, and Level of Difficulty

Independent *t* tests based on the three scores for measurability, equidistance, and level of difficulty for sample 1 showed no statistically significant differences between the groups for each of the features (upper half of Table 4). For the second sample, a statistically significant difference was found for one of the three features, level of difficulty (lower half of Table 4). Bonferroni post hoc comparisons show the web-based group ($M = 2.6$) had statistically significant higher mean difficulty ratings compared to the placebo control group ($M = 2.3$), $t(41) = -2.8$, $p = .008$. No other statistically significant difference was observed between groups.

Given that we did not expect scores on each feature to correlate, we used a conservative alpha level of .0167 (i.e., .05/3, Bonferroni adjustment to control for inflated Type I errors for the set of correlations) for each set of correlations within each study sample. Examination of the intercorrelation matrix (Table 5) based on Pearson correlation coefficient (2-tailed) show that the three qualitative features—measurability, difficulty, and equidistance were not linearly associated with one another in samples 1 and 2.

Interrater Agreement Based on Source of Behavior

Teacher Versus Researcher Supplied

Data from sample 2 allowed us to determine if teachers displayed a differential preference toward showing the best case scenarios of behavior and potentially causing score inflation. A total of 25 out of 29 observations were compared (i.e., four scores were missing because the teacher did not supply a tape of one of the skills to observe). A paired *t* test indicated no statistically significant difference between the mean GAS scores for teacher supplied ($M = -.19$; $SD = .71$) versus researcher supplied ($M = -.40$; $SD = .67$) behavior samples, $t(24) = -1.6$, $p = .11$, $r = .58$, $p = .002$. An ICC of .74, 95% CI (.40, .88) was calculated for average consistency and .58, 95% CI (.25, .79) for single consistency, while ICC for average agreement was .72, 95% CI (.39, .88) and .57, 95% CI (.24, .78) for single agreement.

Video Versus Live Observation

The interrater reliability between ratings based on scores gathered through direct observation and scores coded from

Table 4 Sample 1 and 2 results of GAS features between experimental and control groups

Sample 1	Group							
	Comparison ($n = 17$)		Face ($n = 18$)					
GAS Feature	*M*	*SD*	*M*	*SD*	*t*	*p*	SE_{diff}	η^2
Measurability	2.5	.58	2.6	.51	−.86	.39	.18	.09
Difficulty	2.3	.35	2.1	.45	1.7	.09	.14	.24
Equidistance	2.8	.35	2.8	.30	.4	.69	.11	0

Sample 2	Group									
	Comparison ($n = 15$)		Face ($n = 14$)		Web ($n = 15$)					
GAS feature	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*	*MSE*	η^2_{partial}
Measurability	2.7	.31	2.5	.38	2.4	.34	3.1	.06	.12	.13
Difficulty	2.3	.31	2.5	.38	2.6	.41	3.9	.03	.14	.16[a]
Equidistance	2.8	.28	2.7	.27	2.8	.35	.5	.63	.09	.02

Face face-to-face group, *Web* web-based teacher coaching group

[a] Post hoc analysis using Bonferroni test indicated that that the mean score for the comparison group was statistically significant different from Web. The face-to-face group was not statistically significant different from either of the other two groups

Table 5 Pearson correlations between features in samples 1 and 2

Variable	Measurability	Difficulty	Equidistance
Measurability	–		
Difficulty	−.02 (−.31)	–	
Equidistance	−.02 (.19)	−.04 (−.06)	–

Values in parentheses are from sample 2

videotape of the same behavior sample was conducted for 52% of the observations in sample 2. The primary coder was unaware of group assignment. An ICC of .91, 95% CI (.78, .96) was calculated for average consistency and .83, 95% CI (.64, .92) for single consistency, while ICC for average agreement was .90, 95% CI (.76, .96) and .82, 95% CI (.61, .92) for single agreement. A paired *t* test showed no statistically significant difference in GAS ratings between video supplied mean scores ($M = -.73$; $SD = .71$) and mean scores based on live coding ($M = -.90$; $SD = .79$), $t(22) = 1.83$, $p = .08$, $r = .84$, $p < .001$.

Discussion

The purpose of this study was to establish and evaluate protocols designed to ameliorate concerns of comparability of GAS descriptions across individuals and groups and reliability of GAS scores based on varying levels of data sources. Results from our analyses of two different samples indicate that the assessment of features of measurability, equidistance, and level of difficulty when using GAS are essential. Although data from sample one revealed no differences in the three dimensions across groups, data analysis from sample two did show a statistically significant difference in the level of difficulty in favor of the control group versus the web-based group. This finding means the web-based group consisted of goals that were judged to be more difficult as a group compared to the control group and suggests that researchers might consider including level of difficulty ratings as a covariate or moderator in their group comparisons based on GAS ratings. In general, in order for GAS scores to be able to be treated as equivalent across the groups, it is important that these three criterion measure scores be calculated for each study. Ideally, they should be determined a priori and corrections made if necessary. At the very least, they can be calculated post hoc and if they are not equivalent, then researchers might consider including any one of these three criterion measure scores as covariates or moderators in data analysis following the primary analyses without adjustments, depending on how each criterion relates to GAS scores.

Our question regarding teacher bias in the selection of behavior samples for scoring by the researcher indicated that teachers do not select unrepresentative behavior samples for coding, and concerns that they are conducting multiple observations to select a more favorable outcome are therefore unsupported. This finding is important because it helps validate the approach of teacher selected behavior samples for outcome monitoring. The use of videotapes and other means of technology as alternatives to live data collection will help give teachers, consultants, and

school administrators' confidence in the selection and use of such means. It also reduces the need for resources and travel to schools and classroom. Finally, it lends ecological validity to the approach as teachers made many comments throughout the study of the ease of use of the mini camcorders used and the preference for making their own tapes during the child's natural learning routines, rather than setting up situations that often required children to demonstrate skills during nonroutine times (Jung et al. 2011).

Lastly, analysis of the interrater reliability of GAS scores based on different sources of behavior samples suggests that overall scores are able to be reliably coded. This finding is consistent with previous research (Cytrynbaum et al. 1979; Malec 1999; Schlosser 2004; Shefler et al. 2001; Stolee et al. 1999), indicating that regardless of whether the samples of behavior come from live observations or videotape, ratings are exchangeable for these two methods. However, Schlosser (2004) recommends that reliability of GAS scores be calculated on a case-by-case basis or, in other words, reliability should not be assumed and instead be tested for each study sample. Otherwise, "referencing the reliability coefficients from prior studies as the sole warrant for presuming the score integrity of entirely new data" (p. 512) can create a problem known as "reliability induction" (Vacha-Hasse et al. 2000). Results are mixed on the reliability (comparability) of GAS scores when scales are created by different judges for the same individual (Shefler et al. 2001; Steenbeck et al. 2010). That is, there may be some unreliability in the creation of consistent GAS scales, but once created they tend to be rated similarly. Smith (1994) and Schlosser (2004) concluded, however, that identical goal development should not be a required component of reliability for goal attainment scaling because different goals could be generated from the same problem area by various goal developers.

Recommendations

In summary, we offer several recommendations for creating and applying reliable and valid GAS templates. First, it is important to practice writing GAS templates prior to conducting the study and to use a standardized and systematic approach in writing the GAS goals. We have provided some instructions developed for our studies along with several examples and detailed descriptions (Ruble et al. 2012). To test whether the GAS goals are equivalent between groups, identify naïve coders and have them evaluate the GAS goals using the three dimensions described in this study.

Second, generate GAS objectives prior to onset of the intervention. If the study is longitudinal, it is vital that the goals that are established at the onset of the study remain the same goals that are analyzed at the end of the study for valid between group comparisons. The intervention plans may be adapted or modified according to responsiveness to the intervention, but the goals should remain the same. Third, test whether GAS goals are equivalent using the three dimensions we described and tested. Ideally, this evaluation of equivalence should be conducted prior to implementation of the intervention and any corrections to the GAS goals made at that time. If, however, goal equivalence is evaluated after the onset of intervention, analysis of covariance can be applied if differences between any of the dimensions are observed between groups. Finally, it is necessary to evaluate the assumptions of equivalency and interrater reliability for each study. Our data indicated that even though we were able to reliably code data from the GAS templates, creating equivalent goals may be difficult. Even with strong protocols in place in helping to generate high quality goal descriptions, differences can occur which suggest the need both for a priori carefulness in crafting goals and for post hoc verification of equivalence.

Although several evidence based practices are now available for children with autism, there are still many areas where our knowledge is very limited. Moreover, even when practices are available, there are still problems in how to implement and evaluate the effectiveness of research supported interventions in natural environments. Regardless, a key problem is developing feasible means for evaluating progress, especially when interventions are individually tailored. Based on our findings, goal attainment scaling appears to be a promising idiographic approach for measuring intervention effectiveness.

Acknowledgments This work was supported by Grant Numbers R34MH073071 and 1RC1MH089760 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health. We are grateful to Nancy Dalrymple, co-developer of the COMPASS framework and to the teachers, families, and children who generously donated their time and effort. We wish to thank Rachel Aiello, Jessie Birdwhistell, Ryan Johnson, and Jennifer Hoffman for data coding.

References

Carrow-Woolfolk, E. (1995). *OWLS: Oral and written language scales*. Circle Pines, MN: American Guidance Service, Inc.

Cytrynbaum, S., Ginath, Y., Birdwell, J., & Brandt, L. (1979). Goal attainment scaling: A critical review. *Evaluation Quarterly, 3*(1), 5–40. doi:10.1177/0193841X7900300102.

Elliott, C. D. (1990). *Differential ability scales*. New York: Harcourt Brace Jovanovitch, The Psychological Corporation.

Individuals with Disabilities Education Act. (2004). Retrieved from http://idea.ed.gov/explore/view/p/%2Croot%2C.

Jung, L., Ruble, L., Johnson, R. & McGrew, J. (2011). Web-based coaching of teachers implementing autism interventions (manuscript submitted for publication).

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*, 443–453. doi:10.1007/BF01530764.

Kiresuk, T. J., Smith, A., & Cardillo, J. E. (1994). *Goal attainment scaling: Applications, theory, and measurement*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr, Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule–generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.

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.

Malec, J. F. (1999). Goal attainment scaling in rehabilitation. *Neuropsychological Rehabilitation, 9*(3–4), 253–275. doi:10.1080/096020199389365.

National Research Council. (2001). *Educating children with autism*. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2008).

Oren, T., & Ogletree, B. T. (2000). Program evaluation in classrooms for students with autism: Student outcomes and program processes. *Focus on Autism and Other Developmental Disabilities, 15*, 170–175. doi:10.1177/108835760001500308.

Ottenbacher, K. J., & Cusick, A. (1993). Discriminative versus evaluative assessment: Some observations on goal attainment scaling. *American Journal of Occupational Therapy, 47*, 349–354.

Quenemoen, R., Thurlow, M., Moen, R., Thompson, S. & Morse, A. B. (2003). *Progress monitoring in an inclusive standards-based assessment and accountability system* (Synthesis Report 53). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved Dec 21, 2011, http://education.umn.edu/NCEO/OnlinePubs/Synthesis53.html.

Robins, D. L., Fein, D., Barton, M. L., & Green, J. A. (2001). The modified checklist for autism in toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 31*(2), 131–144.

Ruble, L. A., & Dalrymple, N. J. (2002). COMPASS: A parent-teacher collaborative model for students with autism. *Focus on Autism and Other Developmental Disabilities, 17*(2), 76–83.

Ruble, L. A., Dalrymple, N. J., & McGrew, J. H. (2010a). The effects of consultation on individualized education program outcomes for young children with autism: The collaborative model for promoting competence and success. *Journal of Early Intervention, 32*, 286–301. doi:10.1177/1053815110382973.

Ruble, L., Dalrymple, N., McGrew, J., & Jung, L. (2010b). Examining the quality of IEPs for young children with autism. *Journal of Autism and Developmental Disorders, 40*, 1459–1470. doi:10.1007/s10803-010-1003-1.

Ruble, L., McGrew, J., & Toland, M. (2011). *Comparison of web-based vs face-to-face coaching. Randomized study of teacher training in autism*. Washington, DC: Poster session presented at the American Psychological Association.

Ruble, L., Dalrymple, N. J., & McGrew, J. H. (2012). *Collaborative model for promoting competence and success of students with autism*. NY: Springer.

Rutter, M., Bailey, A., & Lord, C. (2004). *Social communication questionnaire*. Los Angeles, CA: Western Psychological Services.

Schlosser, R. W. (2004). Goal attainment scaling as a clinical measurement technique in communication disorders: A critical review. *Journal of Communication Disorders, 37*, 217–239. doi:10.1016/j.jcomdis.2003.09.003.

Shefler, G., Canetti, L., & Wiseman, H. (2001). Psychometric properties of goal-attainment scaling in the assessment of Mann's time-limited psychotherapy. *Journal of Clinical Psychology, 57*, 971–979.

Sheridan, S. M., Clarke, B. L., Knoche, L. L., & Edwards, C. P. (2006). The effects of conjoint behavioral consultation in early childhood settings. *Early Education and Development, 17*, 593–617. doi:10.1207/s15566935eed1704_5.

Shuster, S. K., Fitzgerald, N., Shelton, G., Barber, P., & Desch, S. (1984). Goal attainment scaling with moderately and severely handicapped preschool children. *Journal of Early Intervention, 8*, 26–37. doi:10.1177/105381518400800104.

Sladeczek, I. E., Elliott, S. N., Kratochwill, T. R., Robertson Mjaanes, S., & Stoiber, K. C. (2001). Application of goal attainment scaling to a conjoint behavioral consultation case. *Journal of Educational and Psychological Consultation, 12*, 45–58. doi:10.1207/S1532768XJEPC1201_03.

Smith, A. (1994). Introduction and overview. In T. Kiresuk, A. Smith, & J. Cardillo (Eds.), *Goal attainment scaling: Applications, theory, and measurement* (pp. 1–14). London: Erlbaum.

Sparrow, S., Cicchetti, D., & Balla, D. (2005). *Vineland adaptive behavior scales* (2nd ed.). Vineland-II: AGS Publishing.

Steenbeck, D., Ketelaar, M., Lindeman, E., Galama, K., & Gorter, J. (2010). Interrater reliability of goal attainment saling in rehabilitation of children with cerebral palsy. *Archives of Physical Medicine and Rehabilitation, 91*, 429–435.

Stolee, P., Zaza, C., Pedlar, A., & Myers, A. M. (1999). Clinical experience with goal attainment scaling in geriatric care. *Journal of Aging and Health, 11*(1), 96–124. doi:10.1177/089826439901100106.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509–522. doi:10.1177/00131640021970682.