American Journal of Evaluation

http://aje.sagepub.com

A Reliability Analysis of Goal Attainment Scaling (GAS) Weights

Stephen M. Marson, Guo Wei and Deborah Wasserman

American Journal of Evaluation 2009; 30; 203
DOI: 10.1177/1098214009334676

The online version of this article can be found at:
http://aje.sagepub.com/cgi/content/abstract/30/2/203

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
American Evaluation Association

American Journal of Evaluation
Volume 30 Number 2
June 2009 203-216

10.1177/1098214009334676
http://aje.sagepub.com
hosted at
http://online.sagepub.com

A Reliability Analysis of Goal Attainment Scaling (GAS) Weights

Stephen M.Marson
Guo Wei
University of North Carolina at Pembroke
Deborah Wasserman
The Ohio State University Center for Family Research

Goal attainment scaling (GAS) has been considered to be one of the most versatile and appealing evaluation protocols available for human services. Aspects of the protocol that make the method so appealing to practitioners—that is, collaboratively working with individual clients to identify and assign weights to goals they will work to achieve—have produced critical psychometric challenges that have threatened the method's acceptance by funders and researchers. This interrater reliability study of weighted goals contributes to their psychometric validation and therefore to the continued use of a methodology so attractive to practitioners. The subjective clinical impressions of 43 students trained in using GAS has statistically significant scorer reliability. These findings suggest that use of GAS composite scores (weight times the problem level) is a reliable tool and therefore not a reason to discourage the use of GAS as a means for monitoring a client's progress over time.

Keywords: *Goal attainment scaling (GAS); program evaluation; interrater reliability; clinical impressions*

Introduction

Utilization has been and continues to be of paramount concern to program evaluation theorists and practitioners. Within that realm, evaluators seek methodologies that engage practitioners by providing information relevant to conducting and improving their daily practice while simultaneously satisfying requirements for accountability. One such method has been goal attainment scaling (GAS), considered by some evaluators and practitioners to be one of the most versatile protocols available for human services (Alter & Evens, 1990). However, methodological constructs that underlie this widely appealing and engaging versatility—that is, collaboratively working with individual clients to identify and assign weights to goals they will work to achieve—have produced critical psychometric challenges that have threatened the method's acceptance by funders and researchers. Many, but not all of those challenges have been satisfactorily addressed in the literature (Schlosser, 2004). One of the unaddressed criticisms relates to the reliability of weighting goals, an aspect of the methodology that

Authors' Note: Stephen M. Marson, Social Work Department, University of North Carolina at Pembroke, Pembroke, NC 28372-1510; e-mail: smarson@nc.rr.com.

practitioners have found to be particularly useful. This study contributes to the continued use of weighted goals by testing the interrater reliability of assigning them.

GAS is an evaluation methodology which, similar to other methodologies for measuring achievement (e.g., academic achievement scores), can be used for both individual assessment and in aggregate to evaluate the program that employs it. The advantage to GAS, however, is that, within a single study, the assessment can be applied across an infinite range of cultures, age groups, and interventions This versatility is made possible through a collaborative relationship between practitioner and client who together identify a unique, customized set of goals and then weight those goals according to their relative contribution to overall achievement. Progress toward these weighted goals can then be translated into a single composite score that can in turn be used for more sophisticated statistical analysis and accountability purposes.

Kiresuk and Sherman (1968) introduced the methodology in 1968 when there were very few evaluative tools designed for mental health practitioners. Since that time, despite the development of more specific outcome measurements, use of GAS has expanded beyond mental health services. Because of its adaptability across micro, mezzo, and macro client systems, and because of its secondary use as a communication tool between program provider and participant, GAS has been encouraged as a method of choice across a wide array of disciplines (Emmerson & Neely, 1988). Beyond mental health, for which the method was developed (Kiresuk & Sherman, 1968; Kiresuk, Smith, & Cardillo, 1994; Maher & Barbrack, 1984; Woodward, Santa-Barbara, Levin, & Epstein, 1978), this array includes social work practice (Rock, 1987), interventions with children (Dreiling & Bundy, 2003; Holroyd & Goldenberg, 1978; Mailloux et al., 2007; Simeonsson, Huntington, & Short, 1982; Young & Chesson, 1997), rehabilitation services (Hurn, Kneebone, & Cropley, 2006), various aspects of education (Glover, Burns, & Stanley, 1994; Roach & Elliott, 2005; Schmidt, Haugaard, & Timmons, 1986), students' academic social skills (Roach & Elliott, 2005), and working with the elderly (Hartman, Borrie, Davison, & Stolee, 1997) and with sex offenders (Hogue, 1994) along with interventions in general (Becker, Stuifbergen, Rogers, & Timmerman, 2000). In his comprehensive critical review of the method, Schlosser (2004) recognized this wide use of GAS across so many disciplines as an indicator of the valuable role it is fulfilling. Today, virtually every textbook that addresses micro intervention also includes a section on how a client's process and outcome can be evaluated by employing GAS.

Essentially, GAS involves four steps. First, in conjunction with a client, the practitioner identifies critical issues that reflect that individual's issues or challenges the program may be able to affect. For example, in the microlevel case study to be presented as central to this article, client issues and challenges involved alcoholism, sexual identity, job performance, and paranoia. A mezzo level evaluation might focus on family functioning or support networks, whereas a macrolevel evaluation might involve public policy changes or voter turnout. The flexibility of identifying issues and expectations around those issues is boundless. In the second step, based on the practitioner's clinical impression, each issue is given a weight—the more problematic the issue, the greater the weight. The third step is, for each challenge or issue, the client and practitioner come to consensus on an expected outcome and to operationalize progress toward those outcomes. Potential progress is commonly defined by specifying each level of an ordinal scale with a range such as the one outlined in Table 1. Together, client and clinician agree on how each level will be described. Finally, the clinician creates a composite score.

For many direct-service providers, the subjective, collaborative aspect of this methodology has helped to alleviate the evaluation anxiety and resistance often encountered with less participatory approaches (Donaldson, Gooler, & Scriven, 2002). Moreover, many of these

Table 1
Typical Levels of a Goal Attainment Scale (to be Further Defined Between Clinician and Client)

Indicator	Value
Much less than expected outcome	−2
Less than expected outcome	−1
Expected outcome	0
More than expected outcome	+1
Much more than expected outcome	+2

practitioners have found the process of weighting the goals to be particularly helpful and the consequent results more meaningful. A summary of GAS attributes includes five major qualities as follows: adaptability, versatility, mutuality, comparability, measurement over time, and the quantitative/qualitative continuum, and because of each of these qualities, utility.

Adaptability

Although the original intent of GAS was to be employed on the microlevel in mental health settings, as has been shown, it has expanded far beyond mental health. Moreover, GAS has also been used successfully on the mezzo (i.e., groups, families) and macro (i.e., community) levels (Alter & Evens, 1990).

Versatility

GAS includes elements of both process and outcome evaluation. It documents types of goals set, the priorities given to those goals, and progress made toward achieving them. Because of this versatility, practitioners can employ GAS to assess process, while concurrently producing data for outcome assessment and accountability, often of more interest to researchers and funders.

Mutuality

Because GAS engages clients in the identification of goals, it is an evaluation method that simultaneously functions as a therapeutic tool. Used optimally, GAS scales are constructed in concert with the client, systemically including client participation in the development of a treatment plan. The clinical term for this participation is *mutuality*. Even after the initial goal selection process, GAS helps discipline the practitioner and client to work closely with each other to achieve these goals.

Comparability

Because scores are customized for each individual participant, GAS allows for comparison across populations, interventions, and subfields (Schlosser, 2004).

Measurement Over Time

GAS measures change over time without risking test–retest threat to internal validity.

Quantitative/Qualitative Continuum

Although producing quantifiable results, GAS goal identification and achievement is closely linked to services so as to be meaningful to practitioners in a way often reserved for more qualitative methodologies.

Utility

Overall, and for many of the reasons listed above, practitioners experience GAS not only as an evaluation system but also a system that functions as a backbone of their daily practice interaction with the people they serve.

Despite clear advantages to GAS as an evaluation method, there are critical and difficult impediments to its use. Schlosser's (2004) review lists these critical areas as involving content and construct validity, reliability, goal scaling, computation and statistical analyses, and rater selection. Schlosser also reviews literature that provides solutions to each of these concerns. However, in his description of the reliability concern, he notes that available information is thin, although what exists is "promising" (p. 228). But reliability information that exists, even in the years since his review, primarily addresses the accuracy of assigning specific scores within a selected goal, not the accuracy of selecting the goal, never mind the weight assigned to it.

Using weights to qualify the relative importance of selected goals—a tool that is key to the method's clinical utility and attractiveness to practitioners—has been shown to be particularly problematic. In fact, in the major text on GAS (Kiresuk, Smith, et al., 1994), Cardillo and Smith strongly recommend against weighting because "applying numerical weights to the scales increases the complexity of the goal-setting task, greatly complicates the calculation of summary GAS scores" (p. 192) and the potentially poor reliability of judging the importance of the scores.

Scriven (1981) addresses the issue of summarizing various evaluative domains into a single evaluative conclusion—the ultimate intention of the GAS composite score. Scriven describes this effort of "weighing and summing" as fundamentally problematic, and in his words "*really* [italics are his] tricky" (p. 86). First, he notes the inherent value decisions necessary for weighting one domain over another. Second, he addresses the issue of the metric and how, when weighting, the relative value changes based on the relativity of the math involved. Finally, he explains the "really nasty" (p. 87) problem encountered with the possibility that some dimensions (goals, in the case of GAS) may cluster together with a weight equal to that of another singular dimension. Were it not for the qualities that make GAS so attractive and usable to practitioners, the difficulty of calculating a valid weighted score may be enough to relegate the method, or at least the weighting aspect, to a back shelf, filed under "nice idea." But, selecting and weighting goals are a key aspect of GAS popularity, one which Schlosser (2004) has labeled as the method's "social validity."

Scriven's (1981) points organize some of the fundamental concerns with utilizing weights. The metric issue is part of an ongoing debate about the appropriate use of nonparametric statistics when the intervals between levels of continuous variables are nondescript. In relation to GAS, the issue has been well addressed. Schlosser has reviewed the varied responses to the suggestion by Mackay, Somerville, and Lundie (1996) that nonparametric tests need to be used. However, one of the by-products of treating the data as nonparametric is to forfeit the use of weights and the composite score they help determine. Fortunately for practitioners who find weights useful, not all responding authors have agreed with the need for the nonparametric approach. If composite scores based on weighted goals are to be used, evidence that interrater reliability for assigning weights is necessary. Evidence of interrater reliability also

would address, at least to some degree, Scriven's first concern about valuing and the third about clustering. In each case, the problem is a bit less "tricky" if it is not compounded by error resulting from variance in rater values. The results of the study presented in this article help reduce the concerns for lack of interrater reliability in regard to assigning weights to selected goals, thus helping to streamline the "tricky" questions related to summarizing weighted data.

Methodology

The Client, the Judges, and GAS

Within a social work course, *Understanding Social Research*, students were introduced to GAS as an evaluative tool for human service intervention. The students received a 3-hr lecture/discussion and completed readings from the text and closed reserve. For a homework assignment, they were given notes from a single client's social history and were told that they would be required to develop four GAS scales.

The subjective case notes concerned Mr. K., a 42-year-old male, admitted to an alcoholic treatment center. They included details from interviews with Mr. K., his wife, his three sons, and his employer. A synopsis of the notes can be found in Table 2. In the synopsis, enough detail is preserved so the reader can understand the broad amount of information student judges considered when constructing and weighting treatment goals.

During class, the students worked together to construct four GAS scales based on the shared social history. At the end of the class period, the students constructed the GAS protocol. Table 3 is an example of one student's work. (Note that the digits 1–7 connote weeks of treatment—seven total weeks. Where the digit is in bold, the student assigned the corresponding row's level of attainment.)

At this point, students were asked to imagine that Mr. K. was their client. They were told to act as if they were seeing this client for seven weekly sessions and *independently* establish weights for the four GAS scales by assigning to each a value between 1 (least severity) and 10 (greatest severity). They were told that the weights were to be based on their clinical assessment of the severity of the four issues determined by the class discussion. Emphasis was placed on the directive that students were *not* allowed to work together in establishing the weights. Besides illustrating the GAS scales developed by the class, Table 3 also includes the clinical impression of one student.

Following Table 3 is Figure 1 derived by multiplying the level of attainment at the time of each session by the weight for the respective goal. The data for the "without weights" line simply adds the unweighted attainment levels.

Note that if either line in Figure 1 is included in a client's chart or even shared with the client, the reader would have a quick vision of the client's progress. However, also note that the line calculated with weights provides a more detailed and accurate picture of what occurred, inclusive of the client's setbacks.

The Sample

Selection of judges for this interrater reliability followed the contention by Drake and Jonson-Reid (2008) that researchers must do everything possible to eliminate alternative

Table 2
Synopsis of Case Notes

Notes From Intake

Mr. K. presented for admission to an alcoholic treatment center as a well dressed 42-year-old male. His blood alcohol analysis showed .26 (legally intoxicated for his weight), but he claimed to have had little to drink; his coordination and his mannerisms were that of a sober person. Mrs. K., Mr. K.'s three sons, and his employer were all involved in convincing Mr. K. to freely admit himself to the center.

At the time of admission, Mr. K. was married with three children. Both K's had well-paying positions, with a combined income of US$120,000 per year. Mr. K. was working for a local utility company, while his wife was an executive administrator for a major corporation. The couple had met while they were students at a Midwestern University and were wed upon graduation. Both Mr. and Mrs. K. were outstanding students with majors in Business Administration. Through most of their marriage, there had been little marital conflict.

Also, at the time of admission, Mr. K. was having serious problems with his employer. His direct supervisor, Mr. L., who has had contact with our alcoholic treatment center, indicated that Mr. K.'s job performance had hit "rock bottom" and that he would be fired if improvements were not made. Mr. L. said that Mr. K. had been one of the company's finest employees but that alcohol seemed to have ruined his effectiveness, and that Mr. K.'s admitted himself to treatment as an alternative to terminating his position with the company. Both Mr. K. and his employer expressed recognition that without good job performance, he has no chance of maintaining his employment.

Mr. K.'s Blue Cross/Blue Shield health insurance would cover 100% of all fees connected with detox up to 7 days and 80% beyond 7–30 days with no further coverage for detox. The policy covered 80% of the fees connected with Inpatient Alcoholic Rehabilitation for a period of 60 days and 40% of the following 60 days with no further coverage for Inpatient Alcoholic Rehabilitation. This policy also covered 80% of outpatient treatment for a period of 24 months as long as the outpatient program is an extension to the inpatient treatment. His Blue Cross would not cover outpatient treatment independent of inpatient treatment for alcoholism.

Notes from intake interview with Mr. K.

At the time of admission, Mr. K. indicated that he was "tired of the problems drinking had caused" him and was "frustrated" that he had to "sink down" to the point of being "forced into treatment". Mr. K. indicated he felt that everyone was against him, and this was the real reason for his preoccupation with alcohol.

Mr. K. reported that he had been reared in Indianapolis, Indiana, where he had been an only child. According to Mr. K., his parents had experienced constant marital discord. Mr. K.'s father had been a highly skilled repairman for IBM, a position that required extended business trips to other parts of the world, at times for up to 2 years. On the rare occasions of Mr. K's homecomings, he and his wife would engage "knock-down, drag-out fights—mostly yelling but no real physical contact." Mr. K. remembered his mother saying nasty things about his dad when he was away. However, Mr. K. noted that neither of his parents were heavy drinkers but did drink alcohol on social occasions. He had no recollection of ever seeing them intoxicated. However, he did remember that his paternal grandfather had a severe drinking problem and died of cirrhosis of the liver. Mr. K. left home to go to college at age 17 and graduated at age 21. By his sophomore year in college, his parents were divorced. He has not seen his father since prior to the divorce and does not care to see or hear from him again. He still sees his mother occasionally on holidays; however, he rarely writes or calls her but does send her money. Mr. K. indicated that his mother told him she would never marry again.

Mr. K. spoke openly and with clarity about how he preferred sexual relationships with men over being with his wife. He said he had never known how to deal with this preference. He noted that he might be addicted to alcohol but indicated that any alcohol abuse evolved from his sexual behaviors. He stated that the only manner in which he could have a sexual relationship with his wife was to first become intoxicated. He also admitted to drinking while searching for and maintaining same-sex relationships. He stated that he was sure that his wife was unaware of these same-sex interests and involvements. He indicated that he "knew" he should either tell her and leave his marriage or to give up his extramarital activities and remain with his wife—but he could not imagine taking either step. He stated that if he left his wife, the standard of living to which he had grown accustomed would change significantly and he did not want that to happen. He indicated that his wife was a "good woman" and deserved better than him. However, he said that he felt "indifferent" toward her and was not sure if his attitudes could change. He suspected that his real feelings toward his wife were "drowned with years of drinking."

Mr. K. also revealed another major concern. Prior to his last drinking binge, he started to worry that he may have acquired syphilis. He said that for him, worse than not knowing what to do about the disease was that he was afraid he had transmitted the disease to his wife. Mr. K. indicated he was confused, scared, and desperate.

(continued)

Table 2 (continued)

Notes from intake interview with Mrs. K.

Mrs. K indicated that the problems gradually began after 10 years of marriage. Mr. K. would return from short business trips intoxicated. Gradually, the trips became longer. Initially, she had thought Mr. K was "seeing another woman," but after several talks with him, concluded that there was no other woman in the picture. Mrs. K. indicated that she believed Mr. K.'s whole problem was alcoholism. She also stated that their three sons, Phillip Jr. (age 22), Thomas (age 20), and Steve (age 19) saw little of their parents.

Notes from intake interview with Mr. K.'s Son

Mr. K.'s son, Phillip Jr. had elected not to go to college and was living nearby, selling insurance. He said his brothers were attending school at a college in another state. He said neither he nor his brothers knew their father well and none of them felt any emotional closeness with their father. He said that while growing up, he was teased by his peers because of what he described as his father's "effeminate mannerisms." He also noted that for him and his brothers, these "mannerisms" were more at issue than any concerns with alcoholism. Together they had wondered if their father's alcoholism was related to his sexual identity. In any case, he said, he and his brothers were in agreement that their father was "different" and needed help.

"Impressions" From Intake Assessment

Mr. K. is a 42-year-old male with an admitting diagnosis of dementia associated with alcoholism (291.22). He is confused, anxious, and frustrated, particularly in dealing with issues of sexuality and communication. He projects the image of being aloof to the feelings of his three sons and wife, having little if any, meaningful communication with them. The reaction of his wife and children may have led Mr. K. to a state of paranoia. Most of Mr. K.'s problems seem to evolve around his addiction to alcohol.

Problem List

- Can Mr. K. and/or his insurance pay for treatment?
- Mr. K. feels that everyone is against him
- Mr. K. and his three sons have a difficult time communicating
- Mr. K. has a very high tolerance for alcohol and must go through a period of detoxification
- Mr. K. has received the medical diagnosis of alcoholism
- Mr. K. has stated that he feels "indifferent" to his wife and is not sure what to do
- Mr. K. is experiencing multiple issues related to his sexual identity
- Mr. K. believes that he has syphilis and does not know what to do
- Mr. K. believes that he has transmitted syphilis to his wife and does not know what to do
- Mr. K. will be fired from his job if his performance does not improve

explanations to the null hypothesis that the method of measurement is unreliable. Thus, they recommend that raters be as homogenous as possible. This group of 43 student raters shared the following characteristics:

- all had the exact GAS training;
- no judge had prior GAS training;
- all shared a single theoretical model for social intervention—the Generalist Model;
- all shared the same educational background, including prerequisite courses;
- all shared similar academic training in establishing interventive priorities for clients;
- most were of the same sex (37 females and 6 males); and
- all shared similar socioeconomic status.

Critical to this study was homogeneity in regard to experience and training with GAS and theoretical understanding of social intervention. Although raters shared the similarities described, they differed by race. The sample included the following: 17 African Americans, 14 Native Americans, and 12 Caucasians.

Table 3
Example of One Student's GAS Protocol With Weights

Levels of Predicted Attainment	Alcoholism wt (10)	Sexual Identity wt (9)	Job Performance wt (7)	Paranoia wt (4)
−2, Much less than expected outcome	Uncontrollable desire to drink	Will not consider sharing information with his wife	Has poor job performance	Feeling that everyone is against him
	1 2 3 4 **5 6 7**	**1 2 3 4 5** 6 7	**1 2 3 4 5 6 7**	**1** 2 **3 4** 5 6 7
−1, Less than expected outcome	Has control some of the time		Shows small amount of job performance improvement	Able to control paranoia feelings
	1 2 **3** 4 5 6 7	1 2 3 4 5 **6 7**	1 2 3 4 5 6 7	1 **2** 3 4 5 6 7
0, Expected outcome	Continues to have desires but is able to control most of the time	Considers but is not willing to have a discussion	Maintains job with adequate evaluations	Feeling comfortable with his paranoia
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
+1, More than expected outcome		Willing to do role playing	Job evaluations show improvement	Functions with little feelings of paranoia
	1 2 3 **4** 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
+2, Much more than expected outcome	Able to control desire to drink	Willing to discuss with wife	Excellent job performance evaluations	Shows no paranoia
	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 **5 6 7**

Note: GAS = goal attainment scaling; wt = weight.

Figure 1
Goal Attainment Scaling Scores With and Without Weights for One Student's Evaluation of Mr. K (as Outlined in Table 3)

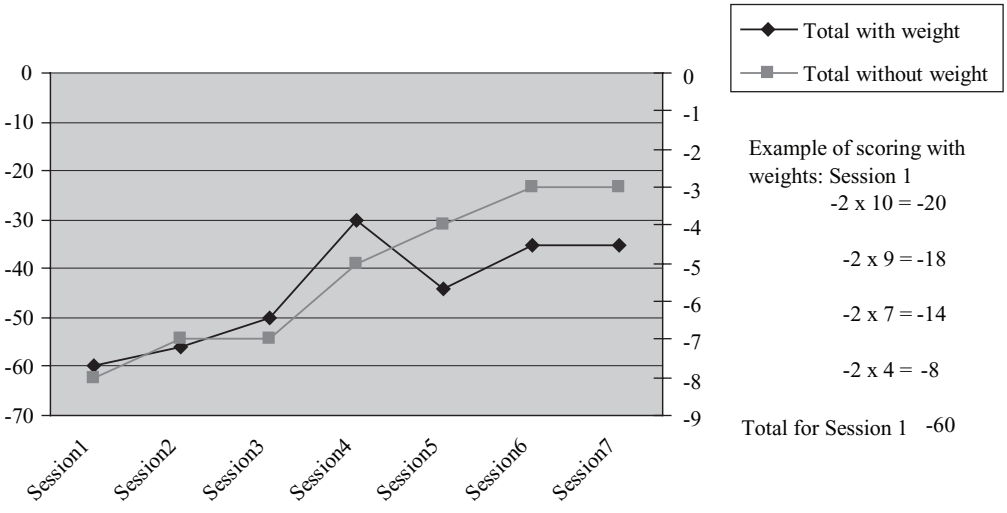

Table 4
Descriptive Statistics: Average Goal Weights and Distributions Among Raters

GAS Issue	*k*	Mean	*SD*	Median	Kurtosis	Skewness
Alcoholism	43	9.72	0.591	10.0	3.13	−2.04
Sexual identity	43	8.56	1.053	39.0	1.72	−0.93
Job performance	43	7.47	1.008	8.0	0.16	−0.49
Paranoia	43	5.79	1.283	6.0	−0.68	−0.09

Note: GAS = goal attainment scaling.

Analysis and Results

As recommended by Gwet (2001), four sequential statistical analyses were used to determine interrater reliability: (a) Kendall's coefficient of concordance *W* (using Friedman's *Q* Chi-square approximation) to establish greater-than-chance similarity between the 43 raters in light of their perception of the differences between four independent goals; (b) a Chi-square test to determine concordance between raters on each individual goal; (c) a Wilcoxon signed rank sum test to measure agreement on the relative distance between goal values within each of the six goal pairs (goal #1–2, 1–3, 1–4, 3–3, 2–4, and 3–4); and (d) intraclass correlation coefficients (ICC) to determine the commonality in the variation between the scores given to each object by the raters. Further descriptions of each of these steps and the consequent results are described below.

Prior to initiating these analyses, we first calculated average weights and distributions across the 43 raters for each goal (Table 4).

Next, Kendall's *W C*oefficient of Concordance was used to eliminate the possibility that rater-to-rater variation was large enough to exclude commonality in raters' evaluation of the four GAS issues (alcoholism, sexual identity, job performance, and paranoia). As per accepted practice (Gibbons, 1993; Noether & Dueker, 1990), Friedman's analysis of variance (ANOVA) was then applied to assure that the commonality would not be because of lack of independence between the goals, that is, that raters' similarity would not be because of the similarity between the goals themselves.

Kendall's *W* coefficient of concordance, used for ranking three or more categorical variables, is the natural extension of Spearman's ρ and Kendall's τ coefficients for measuring association between two variables. According to Siegel & Castellan (1988), the statistic "expresses the degree of association among *k* such variables, that is, the association between *k* sets of rankings . . . " and is "particularly useful in studies of interjudge or intertest reliability" (p. 262).

The formula defining the Kendall's *W* (Gibbons, 1993; Siegel & Castellan, 1988) is as follows:

$$W = \frac{\sum_{i=1}^{N} (\bar{R}_i - \bar{R})^2}{N(N^2 - 1)/12}$$

In this formula, N is the number of issues being ranked, $\bar{R}_i$ is the average of the ranks assigned to the ith object, $\bar{R}$ is the average (grand mean) of the ranks assigned across all objects, and $N(N^2 - 1)/12$ is the maximum possible sum of the squared deviations, that is, the numerator that would occur if there were perfect agreement among the k rankings, and the average rankings were 1, 2, . . . , N. The null hypothesis (Kendall) is that there is no agreement

Table 5
Kendall's *W* Test

Ranks	Mean Rank	Test Statistics	
Alcoholism	3.86	*k*	43
Sexual identity	2.95	Kendall's W[a]	0.83[c]
Job performance	2.07	Chi-square[b]	107.32[d]
Paranoia	1.10	*df*	3

Note: ANOVA = analysis of variance.
a. Kendall's coefficient of concordance.
b. Friedman's ANOVA.
c. $p < .01$.
d. $p < .0001$.

between rankings. Kendall's W is interpreted as a correlation coefficient $0 \leq W \leq 1$. Thus, the closer Kendall's W is to 1, the greater the commonality among the judges. To illustrate Kendall's W, imagine that k judges rank N issues using ranks 1–N with no two items ranked with the same weight: When a perfect agreement is reached for every object, W has a value of 1; when there is no consensus among the judges on the issues so that the average rank for all the GAS issues are exactly the same, W has a value 0. In general, agreement and disagreement are not symmetrical opposites when more than two judges are involved as they may all agree; however, they cannot all disagree completely. For this study, the Kendall's coefficient of concordance W measured the degree of agreement (the extent to which the judges agree) among these k (=43) judges in ranking the N (=4) indicated GAS objects. Hence, if the null hypothesis (there is no agreement) is rejected, the test conclusion is that "the 43 sets of rankings have a high agreement (concordance) although not a perfect one." It should be noted that when tied rankings occur, as was the case in this study, the previous formula for W can be replaced by adjusted rankings (Gwet, 2001; Noether & Dueker, 1990).

Friedman's ANOVA (Gibbons, 1993; Noether & Dueker, 1990; Gwet, 2001) is used to indicate the independence of the ranked categories (the null hypothesis is that the treatment categories are equal). This Q Chi-square statistic further describes Kendall's coefficient of concordance such that the concordance equals the independence divided by the number of judges times one less than the number of categories (in this study, goals) ranked. A high level of concordance with a high level of independence indicates interrater reliability. In contrast, high concordance with low independence would imply that the concordance had more to do with the inability to distinguish between goals than with the similarity between judges. Thus, the value of W can be used to calculate the relation between Friedman's Q and Kendall's W (i.e., $Q = k(N - 1)W$), and the latter is then used to determine whether the k sets of rankings are independent. Because the parameters of this study preclude the typical practice of consulting a table of Q values (W values) for $N = 3$ and $k \leq 15$, or $N = 4$ and $k \leq 8$, the Chi-square approximation $Q = \chi_F^2 = k(N - 1)W$ was employed; that is, under the null hypothesis, the distribution of Q can be approximated by that of the Chi-square χ_F^2 (Gibbons, 1993; Noether & Dueker, 1990).The outcomes of these calculations are used to test the null hypothesis: The k sets of rankings are independent (or to determine whether the data presented sufficient evidence to indicate differences in the distribution of judges' rankings for the four GAS issues).

Notice that the Friedman test (comparing medians of the groups), which generalizes both Sign-Test and the Spearman Rank Correlation Test, with the degrees of freedom $N - 1$, *assumes* one within-participants independent variable (judge) with two or more levels and a dependent variable (GAS objects) that is not interval and normally distributed (but at least

Table 6
Chi-Square Test for Relationship Between Raters for Individual GAS Objects

	Alcoholism	Sexual Identity	Job Performance	Paranoia
Chi-square	40.791	20.140	19.442	12.395
df	2	4	4	5
Asymptotic significant	.000	.000	.001	.030

Note: GAS = goal attainment scaling.

Table 7
Wilcoxon Signed Rank Sum Test for Relationship Between GAS Objects

	Sexual Identity	Job Performance	Paranoia
Alcoholism	318.5**	467.5**	473.0**
Sexual identity		296.0**	467.5**
Job performance			381.5**

Note: GAS = goal attainment scaling.
** Value is significant at a probability of less than .01.

ordinal). In the Friedman test, the two columns are reduced to the Sign-Test. For two rows, the Friedman test is reduced to Spearman Rank Correlation Test. The null hypothesis (Friedman) is that the treatments are equal.

In this data, the calculation of Kendall's W coefficient of concordance involved two steps. The first step was to scale the raw data into the range between 1 and 4 from the original range between 1 and 10, and the second was to apply the Chi-square approximation (χ_F^2; the Friedman test) as described above. As shown in the SAS output illustrated in Table 5, the calculated W was 0.832 (interpreted as a correlation coefficient), which was much larger than the critical value 0.129 at test level $\alpha = .05$. Therefore, the null hypothesis was rejected and the test conclusion was that "the k (=43) sets of rankings have a high agreement (concordance), although not a perfect one." In addition, the calculated χ_F^2 was 107.32 (p value < .0001). Thus, ($W = 107.322/[43 \times (4 - 1)] = 0.832$) indicate that the 43 sets of rankings were not independent at $\alpha = .05$ implying a high association (concordance) of the 43 student judges on their rankings of the four GAS issues. For this calculation, tied observations (which reduce the value of W) were reassigned by the averages—the adjustment method recommended by Siegel and Castellan (1988, p. 264). If tied observations were not considered, the resulting W was 0.820.

Rejection of the null hypothesis with a Kendall's W test does not necessarily signify support for the research hypothesis (high agreement between the judges). Rejecting the null hypothesis only implies that there is a relationship between the two variables (Rater and GAS goal subject). Alternative statistical methods might become necessary to further explain the level of rejection for the null hypothesis and validate the sample because of possible sampling errors (e.g., bias in the data). Hence, for the purpose of gaining greater confidence in the results, additional statistical methods need to be considered.

To assess the relationship between raters on each individual GAS object, a Chi-square test was performed on judges' rankings. The results (shown in Table 6) showed no asymptotic significant independence between the 43 judges ($\alpha = .05$) for each GAS object. In other words, the dependence or concordance between the 43 judges for each of four GAS objects was high.

The third step in the assessment of interrater reliability was to determine significant differences in the relative distances between the weights within each of the six pairs of the four GAS goals

Table 8
Intraclass Correlation Coefficient Between Raters

	Sexual Identity	Job Performance	Paranoia
Alcoholism	−0.66*	−0.14	0.16
Sexual identity		−0.33	0.20
Job performance			−0.32

* When the variations within individual subjects are smaller, but the differences between subject means are larger, the intra-class correlation becomes (positively) larger. In contrast, when the variations within individual subjects are larger, but the differences between subject means are smaller, the intra-class correlation becomes (negatively) larger and this is the effects of other subjects introduced on these subjects.

(Gwet, 2001). The Wilcoxon signed rank sum test is a nonparametric version of a paired-sample *t* test, where the difference between the two variables was not assumed to be interval nor normally distributed. According to this test, significant differences between weights would indicate that raters gave different rankings to the GAS objects. Results of this test demonstrated that concordance existed for the four GAS objects. As illustrated in Table 7, the test values along with the *p* values (close to 0) confirmed the concordance between judges on the four GAS objects.

Finally, interrater reliability needs to reflect commonality in the variation between the scores individual raters assign across the goals (Gwet, 2001). ICC describe homogeneity/consensus in the rankings given by raters (Shrout & Fleiss, 1979). There are several types of ICC, for example, ''the proportion of variance of an observation due to between-subject variability in the true scores'' (Everitt, 1996). The range of the ICC may be between −1 and 1 or between 0.0 and 1.0 depending on the definitions and formulas used. The ICC is typically high when the variation between the scores given to each object by the raters is small. ICC is considered as an improvement over Pearson's *r* and Spearman's ρ because it takes into account of the differences in rankings for individual segments, along with the correlation between raters. The pairwise ICCs between GAS objects are given in Table 8. The ICC between alcoholism and sexual identity is much higher than other pairs (these other pairs have not reached high levels of reliability), and is within a level of interrater reliability that ensures GAS as a reliable measurement. In Royce's (2008) discussion of interrater reliability, the author noted that, ''If the obtained correlation is high (.70 or above), then the researcher has evidence that her rating scale has succeeded in providing a sufficiently reliability measurement'' (p. 150). Within this research, the standard for sufficient reliability was exceeded.

Based on the above statistical analyses, the overall assessment of the rows and columns demonstrate a significant level of common judgment.

Conclusions and Recommendations

This interrater reliability study of weighted goals has responded to concerns that threaten the use of GAS for evaluating human service programs. Results showed that the subjective clinical impressions of 43 students trained in the use of GAS had statistically significant scorer reliability (0.83).

The study demonstrated that GAS weights had an acceptable level of rater reliability (0.832). Many research textbooks use 0.7 (Royce, 2008) or 0.8 (Marlow, 2005; Rubin & Babbie, 2008) as the rule of thumb for acceptable reliability. The existence of scorer reliability is a necessary condition for establishing further research in the area of GAS weights.

Two limitations to this study are important to note. First, raters were all trained uniformly within a short period of time before participating in the study. This condition would be seldom found in the practical application of the methodology. Second, unlike in clinical practice, these ratings were all conducted from a predigested case summary rather than from the students' own case interpretation. For both reasons, our rater agreement could be higher than what would exist in practice. However, both conditions were important to impose in order to isolate and determine the assignment of weights.

When considering the use of weights, two points are important to note. First, to achieve the reliability established in this study, practitioners need training. Excellent training can be found both in (a) Kiresuk, Choate, Cardillo, and Larsen (1994) and (b) at the GAS Web page http://www.marson-and-associates.com/GAS/GAS_index.html. Second, establishing weights for the scales and calculating composite scores from these weights is *not* a time consuming enterprise. In other words, rather than interrupting or reducing face-to-face time with the client—the chief complaint from clinicians about many other evaluation methods GAS enhances it. The process of developing weights with the client assists in establishing rapport.

The interrater reliability reported in this article contributes to evidence that this important and user-friendly evaluation methodology is psychometrically sound. It is hoped that these promising results will encourage other researchers to pursue further documentation of GAS reliability. Further evidence will provide welcome support for a methodology that provides practitioners with a much needed measurement tool—one that satisfies their hunger for evaluation methodology that is relevant, useful, and even beneficial to their daily practice.

References

Alter, C., & Evens, W. (1990). *Evaluating your practice: A guide to self assessment.* New York: Springer.

Becker, H., Stuifbergen, A., Rogers, S., & Timmerman, G. (2000). Goal attainment scaling to measure individual change in intervention studies. *Nursing Research*, *49*, 176-180.

Donaldson, S. I., Gooler, L. E., & Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation*, *23*, 261.

Drake, B., & Jonson-Reid, M. (2008). *Social work research methods.* Boston: Allyn & Bacon.

Dreiling, D. S., & Bundy, A. C. (2003). A comparison of consultative model and direct-indirect intervention with preschoolers. *The American Journal of Occupational Therapy*, *57*, 566-569.

Emmerson, G. J., & Neely, M. A. (1988). Two adaptable, valid, and reliable data-collection measures: Goal attainment scaling and the semantic differential. *Counseling Psychologist*, *16*, 261-271.

Everitt, B. (1996). *Making sense of statistics in psychology.* New York, NY: Oxford University Press.

Gibbons, J. D. (1993). *Nonparametric measures of association.* Newberry Park, CA: Sage.

Glover, S., Burns, J., & Stanley, B. (1994). Goal attainment scaling as a method of monitoring the progress of people with severe learning disabilities. *British Journal of Learning Disabilities*, *22*, 148-150.

Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters.* Gaithersburg, MD: Stataxis.

Hartman, D., Borrie, M. J., Davison, E., & Stolee, P. (1997). Use of goal attainment scaling in a dementia special care unit. *American Journal of Alzheimer's Disease and Other Dementias*, *12*, 111-116.

Hogue, T. E. (1994). Goal attainment scaling: A measure of clinical impact and risk assessment. *Issues in Criminological and Legal Psychology*, *21*, 96-102.

Holroyd, J., & Goldenberg, I. (1978). The use of goal attainment scaling to evaluate a ward treatment program for disturbed children. *Journal of Clinical Psychology*, *34*, 732-739.

Hurn, J., Kneebone, I., & Cropley, M. (2006). Goal setting as an outcome measure: A systematic review. *Clinical Rehabilitation*, *20*, 756-772.

Kiresuk, T. J. Choate, R. O., & Cardillo, J. E. & Larsen, N. (1994). Training in goal attainment scaling. In T. J. Kiresuk, A. Smith, & J. E. Cardillo (Eds.), *Goal attainment scaling: Applications theory, and measurement* (pp. 105-118). Hillsdale, NJ: Lawrence Erlbaum.

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.

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

MacKay, G., Somerville, W., & Lundie, J. (1996). Reflections on goal attainment scaling (GAS): Cautionary notes and proposals for development. *Educational Research*, *38*, 161-172.

Maher, C. A., & Barbrack, C. R. (1984). Evaluating the individual counseling of conduct problem adolescents: The goal attainment scaling method. *Journal of School Psychology*, *22*, 285-297.

Mailloux, Z., May-Benson, T. A., Summers, C. A., Miller, L. J., Brett-Green, B., Burke, J. P., et al. (2007). Goal attainment scaling as a measure of meaningful outcomes for children with sensory integration disorders. *American Journal of Occupational Therapy, 61*, 254-259.

Marlow, C. R. (2005). *Research methods for generalist social work*. Belmont, CA: Brooks/Cole.

Noether, G. E., & Dueker, M. (1990). *Introduction to statistics: The nonparametric way.* New York: Springer.

Roach, A. T., & Elliott, S. N. (2005). Goal attainment scaling: An efficient and effective approach to monitoring student progress. *Teaching Exceptional Children*, *37*, 8-17.

Rock, B. D. (1987). Goal and outcome in social work practice. *Social Work*, *32*, 393-398.

Royce, D. (2008). *Research methods in social work.* Belmont, CA: Brooks/Cole.

Rubin, A. & Babbie, E. (2008). *Research methods for social work*. Belmont, CA: Brooks/Cole.

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.

Schmidt, G., Haugaard, J., & Timmons, G. H. (1986). Guidance program evaluation, goal attainment scaling, and happy thoughts lift winter spirits. *Elementary School Guidance & Counseling*, *20*, 224-230.

Scriven, M.S. (1981). *The logic of evaluation.* Inverness, CA: Edgepress.

Shrout, P., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the social sciences.* New York: McGraw Hill.

Simeonsson, R. J., Huntington, G. S., & Short, R. J. (1982). Individual differences and goals: An approach to the evaluation of child progress. *Topics in Early Childhood Special Education*, *1*, 71-80.

Woodward, C. A., Santa-Barbara, J., Levin, S., & Epstein, N. B. (1978). The role of goal attainment scaling in evaluating family therapy outcomes. *American Journal of Orthopsychiatry*, *48*, 464-476.

Young, A., & Chesson, R. (1997). Goal attainment scaling as a method of measuring clinical outcomes for children with learning disabilities. *The Gerontologist*, *60*, 111-114.