Title

Improving the Precision of the Glasgow Outcome Scale-Extended Using Item Response Theory: A TRACK-TBI Study

Journal

Journal of Neurotrauma, 39(11-12)

ISSN

0897-7151

Authors

Magnus, Brooke E
Balsis, Steve
Giacino, Joseph T
et al.

Publication Date

2022-06-01

DOI

10.1089/neu.2021.0421

Peer reviewed

Journal of Neurotrauma
39:870–878 (June 2022)

DOI: 10.1089/neu.2021.0421

Journal of Neurotrauma

ORIGINAL ARTICLE CLINICAL STUDIES

Improving the Precision of the Glasgow Outcome Scale-Extended Using Item Response Theory: A TRACK-TBI Study

Brooke E. Magnus,[1,*] Steve Balsis,[2] Joseph T. Giacino,[3] Michael A. McCrea,[4] Nancy R. Temkin,[5] John Whyte,[6] Geoffrey T. Manley,[7] Lindsay D. Nelson,[4] and the TRACK-TBI Investigators[**]

Abstract

The Glasgow Outcome Scale-Extended (GOSE) is a functional outcome measure intended to place individuals with traumatic brain injury (TBI) into one of eight broad levels of injury-related disability. This simplicity is not always optimal, particularly when more granular assessment of individuals' injury recovery is desired. The GOSE, however, is customarily assessed using a multi-question interview that contains richer information than is reflected in the GOSE score. Using data from the multi-center Transforming Research and Clinical Knowledge in TBI (TRACK-TBI) study (N = 1544), we used item response theory (IRT) to evaluate whether rescoring the GOSE using IRT, which posits that a continuous latent variable (disability) underlies responses, can yield a more precise index of injury-related functional limitations. We fit IRT models to GOSE interview responses collected at three months post-injury. Each participant's level of functional limitation was estimated from the model (GOSE-IRT) and comparisons were made between IRT-based and standard (GOSE-Ordinal) scores. The IRT scoring resulted in 141 possible scores (vs. 7 GOSE-Ordinal scores in this sample of individuals with GOSE scores ranging between 2 and 8). Moreover, GOSE-IRT scores were significantly more strongly associated with measures of TBI-related symptoms, psychological symptoms, and quality of life. Our findings demonstrate that rescoring the GOSE interview using IRT yields more granular, meaningful measurement of injury-related functional limitations, while adding no additional respondent or examiner burden. This technique may have utility for many applications, such as clinical trials aiming to detect small treatment effects, and small-scale studies that need to maximize statistical efficiency.

Keywords: Glasgow Outcome Scale-Extended; item response theory; outcome measurement; psychometrics

[1]Department of Psychology and Neuroscience, Boston College, Chestnut Hill, Massachusetts, USA.
[2]Department of Psychology, University of Massachusetts Lowell, Lowell, Massachusetts, USA.
[3]Harvard Medical School and Spaulding Rehabilitation Hospital, Charlestown, Massachusetts, USA.
[4]Departments of Neurosurgery and Neurology, Medical College of Wisconsin, Milwaukee, Wisconsin, USA.
[5]University of Washington, Seattle, Washington, USA.
[6]Moss Rehabilitation Research Institute, Elkins Park, Pennsylvania, USA.
[7]University of California, San Francisco, San Francisco, California, USA.
**The TRACK-TBI Investigators may be found at the end of this article.

*Address correspondence to: *Brooke E. Magnus, PhD, Department of Psychology and Neuroscience, Boston College, McGuinn 300, 140 Commonwealth Avenue, Chestnut Hill, Massachusetts, 02467, USA E-mail:* brooke.magnus@bc.edu

Introduction

The Glasgow Outcome Scale Extended (GOSE), along with its previous version, the Glasgow Outcome Scale (GOS), was intended to serve as a straightforward index of functional impairment after traumatic brain injury (TBI) that demarcates patients into distinct outcome categories.[1,2] Subsequently, the GOSE has been used for more diverse research purposes, including translational studies and clinical trials.[2–4]

The GOSE is recommended by the National Institute of Neurological Disorders and Stroke as the outcome measure to be used for major trauma and head injury, and it is the only primary outcome measure that has been accepted by the U.S. Food and Drug Administration for use in TBI research supporting New Drug Application approval.[5] The GOSE has several advantages that explain its popularity, including its efficient and multi-modal administration (e.g., in-person, phone)[2,6] and applicability to diverse injury severities.[7]

Despite widespread use, the GOSE remains subject to criticism.[3,8–10] Some reported limitations include modest reliability,[11–13] misclassification of patients,[11,12,14] lack of sensitivity to small but meaningful change,[5,15] inability to capture more granular individual differences,[9,10] a low ceiling,[3,15,16] and its placement of patients into broadly characterized categories of injury-related disability.[9,17] Notably, many of these limitations center around the instrument's lack of score granularity, which may have contributed to the reduced success of previous clinical trials.[3,5,9] For these reasons, a more granular outcome measure (beyond the 1–8 ordinal GOSE score) may have utility, especially for use in clinical trials.

A semi-structured interview format was introduced in 1998 to help standardize the GOSE and improve the reliability with which GOSE overall scores are assigned.[4] During the interview, the respondent is asked a series of questions across seven domains relating to the patient's ability to participate in daily life, work, and social activities. Injury-related change in each domain is associated with a particular level of disability, and the lowest score implied by the domain responses becomes the overall score. While straightforward, this practice neglects to leverage all available information collected in the interview. For example, it equally weights independence in the home and ability to work.

We intuitively know, however, that these two items may not be equally informative of a patient's disability. Thus, it may be possible to leverage the interviews' question-level data to achieve a more granular index of functional recovery. This has the potential to increase clinical trial efficiency and, importantly, does not increase response burden, because it makes use of data that have already been collected.[9] Item response theory (IRT) is well-suited for this goal.

Our objective was to test the hypothesis that rescoring the GOSE using IRT improves the precision and utility of the instrument for indexing functional recovery. Using data from the multi-center Translating Research and Clinical Knowledge in TBI (TRACK-TBI) study, we addressed two primary questions: (1) Can we use IRT to measure individual differences in TBI severity as assessed by the GOSE when all item-level data are considered; and (2) does IRT scoring improve the measurement properties of the GOSE when compared with the standard GOSE-Ordinal score?

We evaluated measurement properties in terms of the range of disability scores captured by each scoring method, the ''information'' (i.e., inverse of the standard error) that is provided at various levels of impairment, and the strength of the relationships between GOSE scores and criterion measures of constructs closely related to functional recovery (e.g., emotional distress, quality of life).

Method

Study population

We used data at three months post-injury from the TRACK-TBI database (N = 2697 enrolled with TBI), which recruited participants from 18 U.S. level I trauma centers within 24 h of injury and followed them serially thereafter.[18] A final analytic sample of N = 1544 met inclusion criteria for the present study (age >16, not withdrawn before 97 days, and valid GOSE score >1 at three months post-injury). See Table 1 for demographic information for the sample analyzed.

Primary outcome: GOSE

The GOSE is scored on an 8-point ordinal scale, with scores ranging from Death (1) to Upper Good Recovery (8). (For a description of all GOSE levels, see Nelson and colleagues[17].) Only patients classified as Vegetative State (2)* or better were included in this study, because GOSE interview questions were only relevant and scored for TBI survivors. We limited the sample to survivors for two reasons. First, there are no item responses for people who died. Second, we conceptualized the IRT model as placing individuals along a latent continuum of disability, and we do not consider death to be a floor level of disability; rather, we consider it a conceptually distinct category. For these reasons, the IRT scoring approach as described here is limited to survivors.

The TRACK-TBI used the structured interview of the GOSE,[4] which queries individuals with TBI or informants about changes since pre-injury across seven domains. We say ''domain'' to refer to each content area of the interview; domains are assessed with 1–3

*For respondents in a vegetative state, responses indicating maximum impairment were imputed for all subsequent items.

Table 1. Sample Demographics and Injury Characteristics

	Full sample N = 2697	*Final sample[1] N = 1544*
Demographics		
Age, years M (SD)[1]	39.5 (18.7)	40.61 (17.14)
Sex (male)	1859 (68.9%)	1056 (68.4%)
Race		
White	2056 (77.4%)	1185 (77.1%)
Black	429 (16.1%)	252 (16.4%)
Other/unknown	170 (6.5%)	100 (6.5%)
Ethnicity (Hispanic or Latino)	555 (20.8%)	277 (18.0%)
Education, years M (SD)	12.9 (3.5)	13.6 (2.8)
Psychiatric history	575 (21.3%)	357 (23.1%)
Injury characteristics		
Cause of injury		
Motor vehicle/traffic crash	1488 (55.5%)	895 (58.2%)
Fall	762 (28.4%)	404 (26.3%)
Assault/violence	170 (6.3%)	91 (5.9%)
Other/unknown	291 (10.8%)	149 (9.7%)
Highest level of care		
Emergency department	547 (20.3%)	338 (21.9%)
Inpatient floor	940 (34.9%)	543 (35.2%)
Intensive care unit	1210 (44.9%)	663 (42.9%)
TBI severity group		
GCS 13–15 CT-	1256 (50.7%)	742 (50.4%)
GCS 13–15 CT+	721 (29.1%)	452 (30.7%)
GCS 9–12	132 (5.3%)	75 (5.1%)
GCS 3–8	370 (14.9%)	204 (13.8%)
Loss of consciousness[2]	2067 (77.3%)	1303 (84.7%)
Post-traumatic amnesia[2]	1718 (64.3%)	1135 (73.9%)

SD, standard deviation; GCS, Glasgow Coma Scale; CT, computed tomography.

[1]Sample was restricted to ages 17+ with Glasgow Outcome Scale-Extended >1.

[2]Yes or Suspected.

semi-structured interview questions. The measure is scored such that impairment in each domain is associated with a rationally derived disability level, and the lowest score across all domains is used as the overall GOSE (GOSE-Ordinal) score.

As described by Boase and associates,[19] all interview variables were curated through a central review process to minimize missing and inaccurate data and as a tool to maximize consistency in interview practices across sites and examiners. Some questions were either omitted or recoded before IRT analysis; these details can be found in Supplementary Table 1. For the present study, we refer to the recoded questions as items, and it is responses to these items that are used in subsequent analyses.

We scored all items such that higher scores indicated more severe levels of functional disability. The second (Independence at Home) and third (Independence Outside the Home) domains, which each consist of two dichotomous-response items (2a, 2b, 3a, and 3b), were combined into a single item with a 3-point ordinal scale.[+] Responses to two-part questions within the domains of Work, Social and Leisure Activities, and Family and Friendship were collapsed into 3-point or 4-point ordinal items, reflecting all possible outcomes for each domain. The final domain, Return to Normal Life, asks whether any remaining injury-related problems exist (Yes/No).

For the present study, we used GOSE responses that reflected any change in functioning/participation from injury, regardless of whether it occurred because of TBI or peripheral injuries. With the exception of the Work item (which did not apply to 14% of the sample), the percentages of item-level missing data were very small (<0.1%).

[+]Respondents who do not require the assistance of another person for essential at-home activities and are able to shop and travel locally received a score of 0. A score of 1 was assigned to those requiring assistance with shopping or traveling, or those requiring infrequent assistance for in-home activities. Respondents who frequently require the assistance of another person for in-home activities received a score of 2.

Criterion variables

Other injury-related variables and three-month outcome measures were used in criterion validity analyses. The TBI severity was measured by the Glasgow Coma Scale and the presence versus absence of acute intracranial findings on computed tomography scans (CT+/-). Self-reported mild traumatic brain injury (mTBI)-related and psychological symptoms were assessed using the Rivermead Post Concussion Symptoms Questionnaire[20] (RPQ) and the 18-item Brief Symptom Inventory[21,22] (BSI-18) Global Severity Index (GSI). Quality of life was measured with the Quality of Life after Brain Injury Overall Scale[23] (QOLIBRI-OS) and the Satisfaction With Life Scale[24] (SWLS). Missing data percentages were under 10% for all outcome variables.

Statistical analyses

We used a unidimensional IRT model, which assumes that GOSE items reflect a single latent variable (injury-related disability) that sufficiently captures individual differences; the relationship between the latent variable and item responses is modeled with an item response function. As described above, the items submitted for analysis included both a binary item (Return to Normal Life) and several polytomous (3–4 category) items. A common IRT model for binary responses is the two-parameter logistic (2PL), for which each item is characterized by a discrimination (a_j) and a threshold parameter (b_j).

Discrimination describes the relationship between the item and the latent variable, such that the larger the discrimination, the better the item is able to differentiate among levels of disability. A threshold (also called difficulty) reflects the location on the disability continuum where a respondent has 0.5 probability of endorsing the item. Items with higher thresholds are associated with more severe disability.

The graded response model (GRM) is a generalization of the 2PL model that allows for the estimation of discrimination and threshold parameters for items with more than two ordered response categories (e.g., Social Activities). Like the 2PL model, the GRM estimates a

discrimination parameter for each item, (a_j), but unlike the 2PL model, it estimates multiple category-specific threshold parameters for each item. The threshold parameter for category k (b_{jk}) describes the location on TBI-related disability where a respondent has a 0.5 probability of endorsing that category or a more severe one. As is true of the 2PL model, items with higher threshold parameters are associated with more severe disability.

An attractive feature of IRT is that it yields an information function for each item, which indicates how precisely the item can measure individual differences along all points of the TBI-related impairment continuum. Items with larger discrimination parameters provide more information (i.e., greater precision) than those with smaller discriminations; the amount of information an item provides is greatest near its threshold parameter(s).[25]

The sum of the item information functions yields a test information function, which helps summarize how well an entire set of items measures individual differences across levels of disability. Importantly, test information can be thought of as an inverse function of the standard error of measurement from classical test theory, but unlike the standard error of measurement, which is a constant, the test information function varies across the range of the latent variable. Thus, IRT allows us to pinpoint where on the TBI-related impairment continuum the items measure individual differences with the greatest precision.

A further benefit of IRT is that we can compute an IRT scale score and standard error for each respondent. Rather than simply adding responses to obtain a single summed score with the same standard error for everyone in the sample, we can estimate the error around the scale score for each individual's specific response pattern.

We fit a unidimensional IRT model to the item responses at three months, using a hybrid 2PL-GRM model. The model uses full information maximum likelihood estimation, which includes participants with item-level missingness and is robust to missing at random (MAR) data. Fit statistics for this model indicate good fit ($G^2 = 313.18$, $p = 0.04$, RMSEA = 0.01)[#]. An assumption of IRT is that after accounting for the latent variable, no relationships remain among pairs of items—otherwise, items are said to exhibit local dependence.[26]

We evaluated this assumption using χ^2 local dependence statistics after fitting the model, paying particular attention to item pairs with χ^2 values greater than 10. There was some evidence that the Return to Normal Life item shares local dependence with the Family and Friendship ($\chi^2 = 16.6$) and Independence items ($\chi^2 = 15.6$); however, because these statistics were not unusually large and our goal was to preserve all original GOSE items in some form, we kept all items intact. The IRT-estimated functional impairment scale scores (GOSE-IRT, computed as response pattern-based *expected a posteriori* scores or EAPs) and standard errors were computed from the parameter estimates of the final model. All IRT analyses were conducted using IRTPRO.[27]

To evaluate the degree to which IRT-GOSE scores may offer improvement over GOSE-Ordinal scores in terms of criterion validity, we compared dependent robust Spearman or Pearson correlations between GOSE and other measures (injury severity, neuropsychological function, quality of life) using a percentile bootstrap as described by Wilcox.[28] We selected this approach because of the non-normality of the standard GOSE-Ordinal scores. Correlational analyses were performed using the twoDcorR R function developed by Wilcox.[28] At the time of this writing, twoDcorR requires complete cases; analytical samples ranged from 1392 to 1410 for correlational analyses.

Results

The sample was largely male (69%) and white (77%), with an average age of 40 years (standard deviation [SD] = 19). Based on the Glasgow Coma Scale (GCS), 81.1% of participants were classified as having mTBI. Additional sample details can be found in Table 1. Before estimating scale scores, we evaluated the fit of the IRT model and checked whether assumptions were satisfied, as described in the Method section. The model fit well and IRT assumptions were met, with items exhibiting minimal local dependence. The GOSE-Ordinal scores ranged from 2 (most disability) to 8 (least disability); GOSE-IRT scores ranged from -1.11 (least disability) to 2.09 (most disability), resulting in a total of 141 possible scores.

With the exception of GOSE-Ordinal scores of 2 and 8, there was substantial variability in IRT scores at each GOSE-Ordinal score (see Fig. 1). This was particularly true in the middle range of GOSE-Ordinal scores (i.e., 5 and 6). For example, for a GOSE-Ordinal score of 6, GOSE-IRT scores had a mean of 0.16, with a minimum of -0.67 and a maximum 0.76. For the most extreme GOSE-Ordinal scores, there was little to no variability in GOSE-IRT scores.

Items varied in their ability to index the latent construct of functional impairment. The most discriminating item was Independence; the least discriminating one was Family and Friendships (see Table 2). Item and test information curves are shown in Figure 2. Overall, the test was able to capture individual differences with more precision at the higher end of the impairment continuum. Nearly all

[#]We also tested whether it was necessary to estimate a separate discrimination parameter for each item by fitting a version of the model that constrained the discrimination parameters to be equal across the five items. The more complex model fit better than the constrained model, confirming it was helpful to model distinct discrimination parameters across items ($\chi^2[4] = 194.04$, $p < 0.0001$).

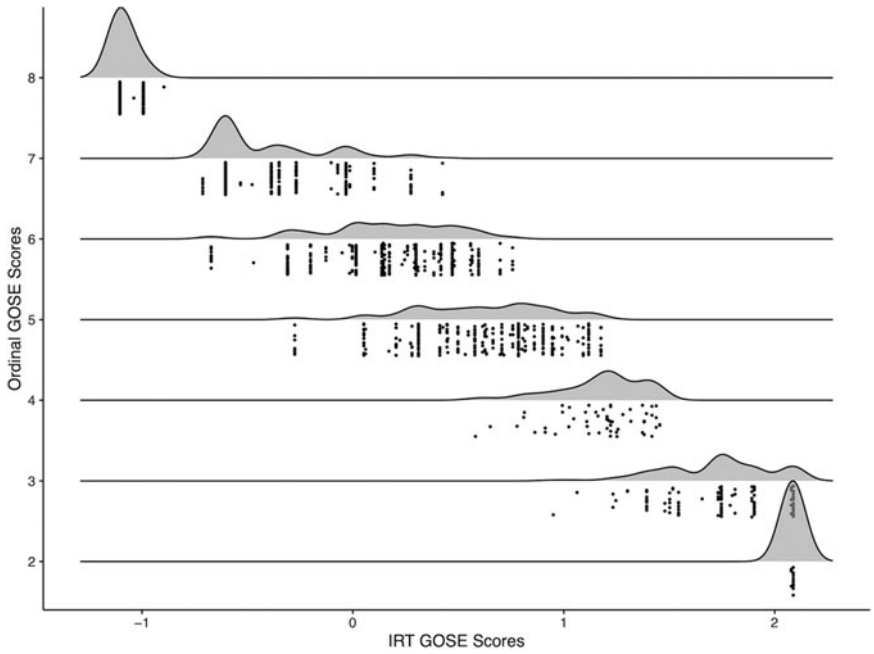

FIG. 1. Scatterplot of respondents' Glasgow Outcome Scale-Extended-item response theory (GOSE-IRT) scores versus standard GOSE-Ordinal scores. Scoring the GOSE through IRT yields a more granular index of functional limitations than the traditional ordinal GOSE, providing a range of IRT scores that correspond to each GOSE score, particularly at GOSE 3–7 levels.

information curves exhibited peaks at impairment levels greater than zero (i.e., above average), which suggests that these items best measure individual differences at moderate and severe levels of disability. While the Returning to Normal Life item is able to differentiate among milder levels of disability, its information curve is very flat, which indicates that this item measures individual differences with much less precision than the items with more pronounced peaks.

The IRT scoring of the GOSE revealed stronger associations with outcome measures than standard ordinal scoring of the GOSE (see Table 3). Whereas both types of scores were moderately correlated with measures of mTBI symptoms, psychological symptoms, and quality of life, GOSE-IRT scores were more strongly associated with worse self-reported symptoms. Notably, these correlations differed in magnitude depending on the type of score that was used: correlations between the GOSE-IRT scores and these self-report measures were significantly larger than those involving GOSE-Ordinal scores, whereas correlations between GOSE scores and markers of injury severity were not statistically different between type of GOSE score.

Table 2. Item Response Theory Parameter Estimates

	a	b_1	b_2	b_3
Independence[1]	5.45	1.24	1.46	–
Work	3.58	0.02	0.58	–
Social & Leisure Activities	3.10	0.20	0.67	1.35
Family & Friendships	1.19	0.85	1.28	2.56
Return to Normal Life	1.46	-0.61	–	–

a, discrimination parameter (i.e., strength of the relationship between the item and the latent disability dimension); *b*, difficulty parameter (i.e., disability/ability level assessed by each variable, for each step from between levels of the variable).

[1]Independence combined Independence in the Home and Independence Outside the home into a 4-level ordinal item because of high correlation between these domains.

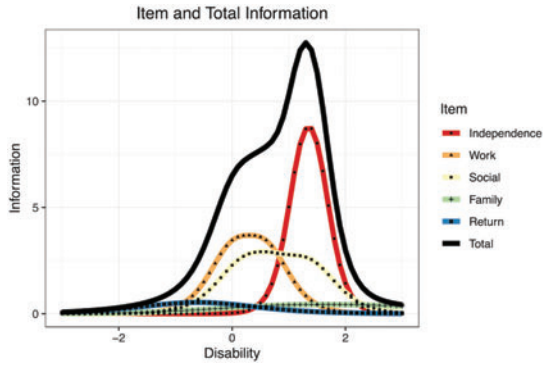

FIG. 2. Item response theory item and total information for domains of the Glasgow Outcome Scale-Extended interview. Color image is available online.

Table 3. Item Response Theory Scoring of the Glasgow Outcome Scale-Extended Reveals Stronger Correlations with Key Self-Report Mild Traumatic Brain Injury and Psychological Symptoms than Standard Scoring of the Glasgow Outcome Scale-Extended

	Spearman's ρ		*Difference*	*Difference*
	GOSE-Ordinal	*GOSE-IRT (reversed)*	*Point estimate [95% CI]*	p
TBI Severity				
Glasgow Coma Scale score	.36	.34	-.01 [-.03, .00]	0.12
Computed tomography findings (+/-)	-.24	-.24	.00 [-.01, .02]	0.83
Psychological symptoms (3 months)				
RPQ	-.61	-.64	.03 [.01, .05]	< 0.001
BSI-18 GSI	-.46	-.50	.04 [.02, .06]	< 0.001
QOLIBRI-OS	.54	.59	.05 [.03, .06]	< 0.001
SWLS	.39	.42	.03 [.01, .05]	0.004

GOSE, Glasgow Outcome Scale-Extended; IRT, item response theory; CI, confidence interval; RPQ, Rivermead Post Concussion Symptoms Questionnaire; BSI-18 GSI, 18-item Brief Symptom Inventory Global Severity Index; QOLIBRI-OS, Quality of Life after Brain Injury Scale-Overall Scale; SWLS, Satisfaction With Life Scale.

For directional consistency, GOSE-IRT scores were reversed for correlational analyses such that lower reflects more disability.

While the differences in self-report measures were modest when considering the overall sample, these differences were more noticeable when stratifying the sample by TBI severity (Tables 4 AND 5). The magnitude of the difference between the correlation coefficients tended to be largest for severe TBI, where differences were as large as 0.1. Finally, when data were stratified by GOSE-Ordinal scores, there were sizable correlations between GOSE-IRT scores and outcome measures (Table 5). With the exception of GOSE scores of 2 and 8 (which have almost zero variability in the corresponding IRT scores), we observed significant, robust correlations between GOSE-IRT scores and psychological outcomes at each GOSE-Ordinal score.

Table 4. Item Response Theory Scoring of the Glasgow Outcome Scale-Extended Stratified by Traumatic Brain Injury Severity Reveals Stronger Correlations with Key Function and Outcome Variables than Standard Scoring of the Glasgow Outcome Scale-Extended

	Spearman's ρ		*Difference*	*Difference*
	GOSE-Ordinal	*GOSE-IRT (reversed)*	*Point estimate [95% CI]*	p
GCS 13–15 CT-				
RPQ	-.67	-.68	.01 [-.01, .03]	0.23
BSI-18 GSI	-.53	-.56	.03 [.01, .05]	0.02
QOLIBRI-OS	.59	.62	.03 [.01, .05]	< 0.001
SWLS	.43	.45	.02 [.00, .04]	0.08
GCS 13–15 CT+				
RPQ	-.63	-.67	.03 [.00, .06]	0.09
BSI-18 GSI	-.50	-.55	.04 [.01, .08]	0.02
QOLIBRI-OS	.52	.57	.05 [.02, .08]	< 0.001
SWLS	.35	.38	.04 [.00, .07]	0.07
GCS 3–12				
RPQ	-.46	-.55	.09 [.03, .15]	0.004
BSI-18 GSI	-.29	-.35	.06 [-.01, .13]	0.10
QOLIBRI-OS	.48	.58	.10 [.04, .16]	< 0.001
SWLS	.36	.42	.06 [-.02, .13]	0.13

GOSE, Glasgow Outcome Scale-Extended; IRT, item response theory; CI, confidence interval; GCS, Glasgow Coma Scale score on admission; CT, computed tomography; RPQ, Rivermead Post Concussion Symptoms Questionnaire; BSI-18 GSI, 18-item Brief Symptom Inventory Global Severity Index; QOLIBRI-OS, Quality of Life after Brain Injury Scale-Overall Scale; SWLS, Satisfaction With Life Scale.

For directional consistency, GOSE-IRT scores were reversed for correlational analyses such that lower reflects more disability.

Discussion

Despite its widespread use, the GOSE has received criticism for its limited scoring precision. The GOSE, however, is typically assessed using a multi-question interview that yields richer information about individual

Table 5. Item Response Theory Scoring of the Glasgow Outcome Scale-Extended Reveals Meaningful Variance at Each Glasgow Outcome Scale-Extended Score (Where Standard Scoring Provides No Useful Variance at Each Integer)

Ordinal GOSE Score	*RPQ*	*BSI-18 GSI*	*QOLIBRI-OS*	*SWLS*
3	-.35	-.42*	.63**	.49*
4	-.38**	-.45**	.40**	.19
5	-.35***	-.29***	.43***	.29***
6	-.29***	-.26***	.30***	.16*
7	-.13*	-.22***	.24***	.13*
8	-.07	-.01	.09	-.02
Combined	-.60***	-.50***	.59***	.44***

GOSE, Glasgow Outcome Scale-Extended; RPQ, Rivermead Post Concussion Symptoms Questionnaire; BSI-18 GSI, 18-item Brief Symptom Inventory Global Severity Index; QOLIBRI-OS, Quality of Life after Brain Injury Scale-Overall Scale; SWLS, Satisfaction With Life Scale; IRT, item response theory.

Correlations reflect Pearson correlations between GOSE-IRT (scaled where higher scores reflect less disability) and each self-report scale denoted by the column headings, within subgroups stratified by GOSE overall score (i.e., GOSE-Ordinal). Ordinal GOSE score of 2 is not shown because there was no variability in IRT GOSE scores.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

outcomes than is reflected by the GOSE overall score it was designed to determine. Whereas the GOSE overall (GOSE-Ordinal) score reflects only the highest level of disability suggested by the interview domains (i.e., minimum domain score), IRT-based GOSE scores leverage additional details recorded during the interview to provide a more granular index of functional disability, which helped to better separate individuals, particularly those in the mid-range of outcome severity. Notably, when compared with the traditional GOSE-Ordinal score, the GOSE-IRT score yielded stronger correlations with self-report measures of mTBI-related symptoms, psychological symptoms, and quality of life, especially for moderate-to-severe levels of disability.

The IRT scoring the GOSE increased its granularity, particularly in cases of incomplete recovery (i.e., GOSE 3–7), because scores in this range can be obtained via many different patterns of responses across the individual items. The GOSE-Ordinal scores of 2 and 8 have more limited associated response patterns, as reflected by the singular GOSE-IRT score at GOSE 2 and minimal variability in GOSE-IRT at GOSE 8 (which was accounted for by how item-level missingness affected our model, rather than the ability of GOSE-IRT to discern levels of mild disability).

While the IRT scoring approach offers less benefit at extreme levels of disability, it is likely to be most helpful for researchers working with TBI samples showing a narrower and more moderate range of disability (i.e., GOSE scores in the 4-6 range). Even more so, the advantage of the IRT approach stands out when stratifying the sample by TBI severity, where IRT scores show moderate correlations with self-reported outcomes at most GOSE-Ordinal scores. Importantly, these correlations would be undetectable if using the traditional scoring approach, because it is assumed that there is no variability in disability among patients who are assigned the same GOSE-Ordinal score.

With the increased variability that IRT scoring creates, the relationships of disability with other outcome variables may be more easily detected, which would likely increase statistical power (and therefore sample size efficiency) for studies using GOSE-IRT score. This is a topic of ongoing research.

In addition to improved granularity, a further benefit of IRT scoring is that the technique takes into consideration how well each item discriminates among levels of disability along the latent variable continuum—that is, items that are more discriminating carry more weight than items that are less discriminating. Items that do not differentiate well among individuals will not have as much leverage in the computation of scores, which is in contrast to how the GOSE is traditionally scored.

It is important to note that the IRT scoring approach developed here introduces no additional administration burden for examiners or patients, because it makes use of the information that is already routinely collected in GOSE interviews. Because we found that IRT scoring performed as well as or better than traditional scoring, we believe that IRT can play a key role in the improved measurement of TBI outcome.

Researchers with access to item-level GOSE data can fit IRT models and estimate item parameters with widely available software, allowing them to calculate scale scores as we have done here. Similarly, accessible tools could be developed that use IRT model parameters such as those established in this study to compute GOSE-IRT from interview responses. We would advocate, however, for additional validation of this model (e.g., to verify measurement invariance across important subgroups and over time) before using any particular set of item parameters as part of a centralized GOSE-IRT scoring system.

In future studies, it would also be useful to compare the two types of scores longitudinally, because the increased variability of IRT scores may increase their sensitivity to change. It may also be informative to empirically determine the IRT cut scores that correspond to commonly used favorable/unfavorable dichotomizations of the GOSE. Moreover, one could examine whether augmenting the item-level GOSE data with items from other related outcome assessments can further improve measurement precision and outcome classification within an IRT framework.[29]

Finally, it is important to note that we only considered GOSE ratings reported because of both the effects of the TBI as well as any concurrent peripheral injuries. Sometimes the GOSE is administered to isolate the TBI-related contribution to impairment; in this case, it would be important to evaluate whether the model holds.

Conclusion

We showed that IRT has the potential to improve the measurement of TBI-related disability and provide a more granular characterization of TBI outcome compared with conventional GOSE scoring. In particular, traditional ordinal scoring of the GOSE results in a small number of possible scores and, consequently, minimal separation of respondents in terms of disability severity. In contrast, IRT scoring the responses routinely collected in the GOSE interview yields more than 100 different possible scores. Importantly, the increased variability of scores reflects meaningful individual differences, because these scores are more highly correlated with other outcome measures of symptoms and quality of life. Improved separation and measurement of individuals along the functional impairment continuum has important implications for precision medicine initiatives and clinical trial outcome measurement.

Acknowledgments

The manuscript's contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

TRACK-TBI Investigators

Neeraj Badjatia, University of Maryland, College Park, MD; Ramon Diaz-Arrastia, University of Pennsylvania Perelman School of Medicine, Phiadelphia, PA; Shankar Gopinath, Baylor College of Medicine, Houston, TX; Ramesh Grandhi, University of Utah, Salt Lake City, UT; Sonia Jain, University of California, San Diego, San Diego, CA; Ruchira M. Jha, Barrow Neurological Institute, Phoenix, AZ; C. Dirk Keene, University of Washington, Seattle, WA; Christine Mac Donald, University of Washington, Seattle, WA; Christopher Madden, UT Southwestern, Dallas, TX; Laura B. Ngwenya, University of Cincinnati, Cincinnati, OH; David Okonkwo, University of Pittsburgh, Pittsburgh, PA; Claudia Robertson, Baylor College of Medicine, Houston, TX; Richard B. Rodgers, Goodman Campbell Brain and Spine, Carmel, IN; David Schnyer, UT Austin, Austin, TX; Andrea Schneider, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; Sabrina R. Taylor, University of California, San Francisco, San Francisco, CA; Abel Torres Espin, University of California, San Francisco, San Francisco, CA; John K. Yue, University of California, San Francisco, San Francisco, CA; Ross Zafonte, Harvard Medical School, Boston, MA.

Funding Information

This secondary data analysis project was supported by NINDS grant # R01 NS110856. The TRACK-TBI study was funded by the U.S. National Institute for Neurological Disorders and Stroke (NINDS) grant # U01 NS1365885.

Author Disclosure Statement

No competing financial interests exist.

Supplementary Material

Supplementary Table S1

References

1. Jennett, B., Bond, M.R. (1975). Assessment of outcome after severe brain damage. Lancet 1, 480–484.
2. Jennett, B., Snoek, J., Bond, M.R., and Brooks, N. (1981). Disability after severe head injury: observations on the use of the Glasgow Outcome Scale. J. Neurol. Neurosurg. Psychiatry 44, 285–293.
3. McMillan, T., Wilson, L., Ponsford, J., Levin, H., Teasdale, G., and Bond, M. (2016). The Glasgow Outcome Scale—40 years of application and refinement. Nat. Rev. Neurol. 12, 477–485.
4. Wilson, J.T., Pettigrew, L.E., and Teasdale, G.M. (1998). Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. J. Neurotrauma 15, 573–585.
5. Yeatts, S.D., Palesch, Y.Y., and Temkin, N. (2018). Biostatistical issues in TBI clinical trials, in: *Handbook of Neuroemergency Clinical Trials*. Elsevier: Cambridge, MA, pps. 167–185.
6. Wilson, J.T.L., Edwards, P., Fiddes, H., Stewart, E., and Teasdale, G. M. (2002). Reliability of postal questionnaires for the Glasgow Outcome Scale. J. Neurotrauma 19, 999–1005.
7. Wilson, J.T.L., Pettigrew, L.E.L., and Teasdale, G.M. (2000). Emotional and cognitive consequences of head injury in relation to the Glasgow Outcome Scale. J. Neurol. Neurosurg. Psychiatry 69, 204–209.
8. Nelson, L.D., Magnus, B.E., Temkin, N., Dikmen, S. and Balsis, S. (2021). Functional Status Examination yields higher measurement precision than the Glasgow Outcome Scale-Extended after moderate-to-severe traumatic brain injury. J. Neurotrauma 38, 3288–3294.
9. Ranson, J., Magnus, B.E., Temkin, N., Dikmen, S., Giacino, J.T., Okonkwo, D.O., Valadka, A.B., Manley, G.T., Nelson, L.D., and TRACK-TBI Investigator. (2019). Diagnosing the GOSE: structural and psychometric properties using Item Response Theory, a TRACK-TBI pilot atudy. J. Neurotrauma 36, 2493–2505.
10. Nelson, L.D., Brett, B.L., Magnus, B.E., Balsis, S., McCrea, M.A., Manley, G.T., Temkin, N., and Dikmen, S. (2020). Functional Status Examination yields higher measurement precision of functional limitations after traumatic injury than the Glasgow Outcome Scale-Extended: a preliminary study. J. Neurotrauma 37, 675–679.
11. Lu, J., Murray, G.D., Steyerberg, E.W., Butcher, I., McHugh, G.S.; Lingsma, H., Mushkudiani, N., Choi, S., Maas, A.I.R., and Marmarou, A. (2008). Effects of Glasgow Outcome Scale misclassification on traumatic brain injury clinical trials. J. Neurotrauma 25, 641–651.
12. Teasdale, G.M., Pettigrew, L.E.L., Wilson, J.T.L., Murray, G.D., and Jennett, B. (1998). Analyzing outcome of treatment of severe head injury: a review and update on advancing the use of the Glasgow Outcome Scale. J. Neurotrauma 15, 587–597.
13. Maas, A.I.R., Braakman, R., Schouten, H.J.A., Minderhoud, J.M., and Zomeren, A.H. (1983). Agreeent between physicians on assessment of outcome following severe head injury. J. Neurosurg. 58, 321–325.
14. Wilson, J.T.L., Slieker, F.J.A., Legrand, V., Murray, G., Stocchetti, N., and Maas, A.I.R. (2007). Observer variation in the assessment of outcome in traumatic brain injury: experience from a multicenter, international randomized clinical trial. Neurosurgery 61, 123–128.
15. Bullock, M.R., Merchant, R.E., Choi, S. C., Gilman, C.B., Kreutzer, J.S., Marmarou, A., and Teasdale, G.M. (2002). Outcome measures for clinical trials in neurotrauma. Neursurg. Focus 13, ECP1.
16. Hudak, A.M., Caesar, R.R., Frol, A.B., Krueger, K., Harper, C.R., Temkin, N.R., Dikmen, S.S., Carille, M., Madden, C., and Diaz-Arrastia, R. (2005). Functional outcome scales in traumatic brain injury: a comparison of the Glasgow Outcome Scale (Extended) and the Functional Status Examination. J. Neurotrauma 22, 1319–1326.
17. Nelson, L.D., Ranson, J., Ferguson, A.R., Giacino, J., Okonkwo, D.O., Valadka, A., Manley, G., and McCrea, M. (2017). Validating multidimensional outcome assessment using the TBI common data elements: an analysis of the TRACK-TBI pilot study sample. J. Neurotrauma 34, 3158–3172.
18. Nelson, L.D., Barber, J.K., Temkin, N.R., Dams-O'Connor, K., Dikmen, S., Giacino, J.T., Kramer, M.D., Levin, H.S., McCrea, M.A., Whyte, J., Bodien, Y.G., Yue, J.K., Manley, G.T., and British Neurosurgical Trainee Research Collaborative (BNTRC). (2021). Validity of the brief test of adult cognition by telephone in level 1 trauma cancer patients six months post-traumatic brain injury: a TRACK-TBI study. J. Neurotrauma 38, 1048–1059.
19. Boase, K., Machamer, J., Temkin, N., Dikmen, S., Wilson, L., Nelson, L.D., Barber, J., Bodien, Y.G., Giacino, J.T., Markowitz, A.J., McCrea, M.A., Satris, G., Stein, M.B., Taylor, S.R., Manley, G.T. and the TRACK-TBI Investigators (2021). Central curation of Glasgow Outcome Scale-Extended data: lessons learned from TRACK-TBI. J. Neurotrauma 38, 2419–2434.
20. King, N., Crawford, S., Wenden, F., Moss, N., and Wade, D. (1995). The Rivermead Post Concussion Symptoms Questionnaire: a measure of symptoms commonly experienced after head injury and its reliability. J. Neurol. 242, 587–592.
21. Derogatis, L., and Melisaratos, N. (1983). The Brief Symptom Inventory: an introductory report. Psychol. Med 13, 595–605.
22. Derogatis, L.R. (2001). *BSI 18, Brief Symptom Inventory 18: Administration, Scoring and Procedure Manual.* NCS Pearson: Minneapolis.
23. von Steinbuchel, N., Wilson, L., Gibbons, H., Hawthorne, G., Hofer, S., Schmidt, S., Bullinger, M., Maas, A., Neugebauer, E., Powell, J., von Wild, K., Zitnay, G., Bakx, W., Christensen, A.L., Koskinen, S., Sarajuuri, J., Formisano, R., Sasse, N., Truelle, J.L. and QOLIBRI Task Force. (2010). Quality of Life after Brain Injury (QOLIBRI): scale development and metric properties. J. Neurotrauma 27, 1167–1185.
24. Diener, E., Emmons, R.A., Larsen, R.J., and Griffen, S. (1985). The Satisfaction with Life Scale. J. Pers. Assess. 49, 71–75.

25. Embretson, S.E., and Reise, S.P. (2000). *Item Response Theory for Psychologists.* Lawrence Erlbaum Associates, Inc.: Mahwah, NJ.
26. Chen, W.H., and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. J. Educ. Behav. Stat. 22, 265–289.
27. Cai, L., Du Toit, S., and Thissen, D. (2011). *IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling.* Vector Psychometric Group, LLc: Seattle, WA.
28. Wilcox, R.R. (2016). Comparing dependent robust correlations. Br. J. Math. Stat. Psychol. 69, 215–224.
29. Whyte, J., Giacino, J.T., Heinemann, A.W., Bodien, Y., Hart, T., Sherer, M., Whiteneck, G.G., Mellick, D., Hammond, F.M., Semik, P., Rosenbaum, A., and Richardson, R.N. (2021). Brain Injury Functional Outcome Measure (BI-FOM): A single instrument capturing the range of recovery in moderate-severe traumatic brain injury. Arch. Phys. Med. Rehabil. 102, 87–96.