The Neurological Outcome Scale for Traumatic Brain Injury (NOS-TBI): II. Reliability and Convergent Validity

Stephen R. McCauley,¹ Elisabeth A. Wilde,² Tara M. Kelly,³ Annie M. Weyand,⁴
Ragini Yallampalli,⁵ Eric J. Waldron,⁵ Claudia Pedroza,⁶ Kathleen P. Schnelle,⁷ Corwin Boake,⁷
Harvey S. Levin,⁷ and Paolo Moretti⁸

Abstract

A standardized measure of neurological dysfunction specifically designed for TBI currently does not exist and the lack of assessment of this domain represents a substantial gap. To address this, the Neurological Outcome Scale for Traumatic Brain Injury (NOS-TBI) was developed for TBI outcomes research through the addition to and modification of items specifically relevant to patients with TBI, based on the National Institutes of Health Stroke Scale. In a sample of 50 participants (mean age = 33.3 years, SD = 12.9) ≤ 18 months (mean = 3.1, SD = 3.2) following moderate ($n = 8$) to severe ($n = 42$) TBI, internal consistency of the NOS-TBI was high (Cronbach's $\alpha = 0.942$). Test-retest reliability also was high ($\rho = 0.97$, $p < 0.0001$), and individual item kappas between independent raters were excellent, ranging from 0.83 to 1.0. Overall inter-rater agreement between independent raters (Kendall's coefficient of concordance) for the NOS-TBI total score was excellent ($W = 0.995$). Convergent validity was demonstrated through significant Spearman rank-order correlations between the NOS-TBI and the concurrently administered Disability Rating Scale ($\rho = 0.75$, $p < 0.0001$), Rancho Los Amigos Scale ($\rho = -0.60$, $p < 0.0001$), Supervision Rating Scale ($\rho = 0.59$, $p < 0.0001$), and the FIM[™] ($\rho = -0.68$, $p < 0.0001$). These results suggest that the NOS-TBI is a reliable and valid measure of neurological functioning in patients with moderate to severe TBI.

Key words: convergent validity; Neurological Outcome Scale for Traumatic Brain Injury; outcome; reliability; traumatic brain injury

Introduction

AS PREVIOUSLY DISCUSSED BY WILDE ET AL. (Wilde et al., 2010a), a standardized measure of neurological functioning in patients with traumatic brain injury (TBI) does not exist, which, coupled with the general lack of assessment of neurological functioning in these patients, has impeded progress in intervention trials for TBI (Narayan et al., 2002). To address this gap, the Neurological Outcome Scale for Traumatic Brain Injury (NOS-TBI) was developed (Wilde et al., 2010a), which measures neurological functioning spe-

cifically for use in patients with TBI, and was based on the well-known and widely-used National Institutes of Health Stroke Scale (NIHSS; Brott et al., 1989; Goldstein et al., 1989; Josephson et al., 2006; Pallicino et al., 1992; Powers, 2001; Schlegel et al., 2003; Spilker et al., 1997; Sun et al., 2006). The NOS-TBI was designed to more appropriately measure common salient post-TBI sequelae having important implications for rehabilitation and outcome. While the studies by Wilde and associates (Wilde et al., 2010a, 2010b) have outlined the feasibility of the NOS-TBI, its use in clinical and research applications, and detailed the instrument's construct validity

¹Physical Medicine and Rehabilitation Alliance of Baylor College of Medicine and the University of Texas–Houston Medical School, and the Departments of Neurology and Pediatrics, ²Physical Medicine and Rehabilitation Alliance of Baylor College of Medicine and the University of Texas–Houston Medical School and the Departments of Neurology and Radiology, Baylor College of Medicine, Houston, Texas.

³University of Texas–Houston Medical School, Houston, Texas.

⁴Baylor College of Medicine, Houston, Texas.

⁵University of Houston–University Park, Houston, Texas.

⁶Center for Clinical Research and Evidence-Based Medicine at the University of Texas Medical School at Houston, Houston, Texas.

⁷Physical Medicine and Rehabilitation Alliance of Baylor College of Medicine and the University of Texas–Houston Medical School, Houston, Texas.

⁸Departments of Neurology and Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas.

(the degree to which an instrument measures an operationalized or underlying theoretical concept or construct), and content validity (the extent to which an instrument's test items adequately cover a representative sample of the domain of interest), this paper details additional critical psychometric properties of the NOS-TBI to demonstrate its applicability to TBI outcomes research.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) outlines technical standards for the construction and evaluation of test instruments. The two pillars of the standards include reliability (the consistency of scores obtained or the extent to which test scores reflect "true" scores, with a minimum of chance or other error variance), and validity (what a test actually measures and the degree to which it does so) of test instruments. Each of these requirements is equally vital for the appropriate use and interpretation of measurements that such test instruments generate. Therefore, the purpose of this study was to investigate the reliability (internal consistency, and test-retest and inter-rater agreement), and validity (specifically, convergent validity) of the NOS-TBI to demonstrate these psychometric properties and to facilitate its potential usefulness in assessing neurological dysfunction in patients with moderate to severe TBI.

Methods

Informed consent was obtained from the participant, a legal authorized representative, or a parent/guardian (for adolescents under 18 years of age) through an informed consent form, and the procedure was approved by the Institutional Review Board of Baylor College of Medicine and its affiliate institutions.

Sample characteristics

Participants in this study are the same as those reported by Wilde and colleagues (Wilde et al., 2010b). All participants were fluent in either English or Spanish, and assessment was performed in the patient's preferred language. Professional interpreters on staff at the rehabilitation facility were involved as necessary to assist in assessment of Spanish monolingual participants. Inclusion criteria included patients between the ages of 15 and 65 years (inclusive) who sustained TBI. Patients were excluded if they had evidence of a penetrating head injury, spinal cord injury, history of a premorbid neurological disorder, major psychiatric disorder (e.g., schizophrenia or bipolar disorder), or if they were >18 months post-TBI. Injury and demographic data are presented in Tables 1 and 2. The typical participant was a non-Hispanic Caucasian male sustaining a severe TBI in a motor vehicle accident with positive findings on computed tomography (CT) imaging performed emergently. Participants were evaluated while undergoing inpatient rehabilitation.

Neurological Outcome Scale for Traumatic Brain Injury

Details of the NOS-TBI scale, and its development, content, and construct validity, have been presented previously (Wilde et al., 2010a, 2010b). To summarize briefly, items from the NIHSS were modified such that the resulting test items were more fine-grained (e.g., quadrant-based testing of motor and sensory functions instead of bilateral testing, and later-

TABLE 1. CATEGORICAL VARIABLES:
DEMOGRAPHICS AND INJURY CHARACTERISTICS

	n	%
Gender		
Male	45	90
Female	5	10
Mechanism of injury		
MVA occupant	36	72
Fall	8	16
Assault	2	4
Bicycle	2	4
Auto-pedestrian	1	2
Other	1	2
Race		
European American	46	92
African-American	2	4
Native American	1	2
Multiracial	1	2
Ethnicity		
Hispanic	11	22
Non-Hispanic	39	78
Occupation		
Professional/technical	4	8
Managerial/clerical	13	26
Craftsperson/skilled labor	14	28
Unemployed, homemaker, student, retired	8	16
Operatives/semi-skilled labor	8	16
Unskilled labor	3	6
Injury severity		
Moderate ^a	8	16
Severe	42	84
Admission CT scan result		
Negative	2	4
Positive	48	96
Loss of consciousness		
No	1	2
Yes	49	98

MVA, motor vehicle accident; TBI traumatic brain injury; CT, computed tomography.

^aParticipants with complicated mild injuries were classified as having a moderate TBI, given the similarity of their neuropsychological performance profiles (Williams et al., 1990).

alized assessment of visual fields, facial paresis, and limb ataxia replacing unidimensional tests), and items assessing neurological functions specifically relevant to TBI (pupillary response, olfaction, bilateral hearing assessment, and gait ataxia) were added.

Outcome measures used for validation

Instruments selected for validation of the NOS-TBI were selected on the basis of demonstrated acceptable psychometric properties of reliability and validity, and their common use in acute and/or post-acute TBI rehabilitation. The Disability Rating Scale (DRS) is a frequently used, well-validated rating scale of disability following brain injury that measures the patient's level of functioning from coma to community re-entry (Gouvier et al., 1987; Hall et al., 1985, 2001; McCauley et al., 2001a; Rappaport et al., 1981). The Supervision Rating Scale (SRS) measures the level of supervision that a patient with TBI actually receives from staff or other caregivers on a 13-level rating scale (Boake, 1996). The Rancho Los Amigos

TABLE 2. CONTINUOUS VARIABLES: DEMOGRAPHICS, INJURY, RELIABILITY, AND VALIDITY MEASURES

Variable	Mean	SD	Median	Minimum	Maximum
Age at assessment	33.3	12.9	29.2	15	65
Post-resuscitation GCS score	6.0	3.2	6	3	15
Education (years)	13.0	2.3	12	9	18
Time post-injury (months)	2.9	2.4	2	0.4	11.5
NOS-TBI test-retest interval (days)	1.3	0.9	1	0	6
NOS-TBI to neurological exam interval (days)	1.5	1.6	1	0	8
DRS	8.2	5.0	6	1	22
SRS	7.8	2.3	7	4	12
RLAS	7.1	1.5	8	2	8
FIM ^a (<i>n</i> = 46)	69.3	34.9	79	18	119

^aThe Functional Independence Measure (FIM) scores reported here were those closest to the administration of the NOS-TBI (conducted either at inpatient rehabilitation admission or discharge time points). FIM scores obtained >1 month before or after the administration of the NOS-TBI were excluded from the analyses.

GCS, Glasgow Coma Scale; NOS-TBI, Neurological Outcome Scale for Traumatic Brain Injury; DRS, Disability Rating Scale; SRS, Supervision Rating Scale; RLAS, Rancho Los Amigos Scale; SD, standard deviation.

Scale (RLAS; also referred to as “Rancho” or the “Levels of Cognitive Functioning Scale”) is an 8-level rating scale of a patient’s overall level of consciousness and cognitive and behavioral functioning that is commonly used to classify patients in acute and post-acute rehabilitation settings (Gouvier et al., 1987; Hagen et al., 1979). The Functional Independence Measure (FIMTM) instrument is a widely-used, 7-level, 18-item ordinal scale measuring both physical and cognitive disability during inpatient rehabilitation with demonstrated reliability and validity (Granger et al., 1990, 1993; Hamilton et al., 1994; Heinemann et al., 1997; Keith et al., 1987; Linacre et al., 1994). Higher scores on the DRS and SRS indicate greater functional impairment, whereas the inverse is true for the RLAS and FIM.

Testing procedure

Participants currently involved in an inpatient rehabilitation program were screened for eligibility. After chart screenings and approval from the patient’s attending physician were obtained, participants were invited to undergo assessment using the NOS-TBI. Testing occurred in the participant’s hospital room. During the initial assessment, Rater 1 (E.A.W.) performed the NOS-TBI, while Rater 2 observed. Scores on the NOS-TBI were recorded by both raters without consultation between them. This procedure was used to reduce concerns regarding minimization of participant burden and maintenance of independence of ratings. As part of the validation of the NOS-TBI, participants were assessed on a second occasion, and also received a full neurological examination (Wilde et al., 2010b). Having the initial assessment performed by Rater 1 and observed (and independently scored) by Rater 2 reduced the participant’s involvement from four separate assessments to three. Assessment of excessively somnolent and/or physically weak patients as a result of undergoing rehabilitative therapies was deferred until they were more alert and rested. All test items were presented in the same order for each participant.

Statistical analysis

All analyses were conducted using SAS software for Windows, version 9.2 (SAS Institute Inc., 2008). Statistical

significance was defined as $\alpha = 0.05$ for all analyses unless otherwise specified. Spearman rank-order correlations and intraclass correlations were performed. Data were screened for entry errors and necessary corrections were made. No participant had missing data for any item on the NOS-TBI.

Results

Internal consistency

Internal consistency is defined as the extent to which a set of test items of an instrument may be treated as measuring a single latent variable or construct (Anastasi and Urbina, 1997; Nunnally and Bernstein, 1994). Internal consistency of the NOS-TBI was measured through Cronbach’s coefficient alpha (range 0–1.0), which is preferable to the Kuder-Richardson 20 formula, as the NOS-TBI includes item responses that are not dichotomous. The alpha for all items of the NOS-TBI (as scored by Rater 1, E.A.W.) was 0.942, which is considered a satisfactory level of internal consistency, as it exceeds the widely-accepted 0.70 criterion (Nunnally and Bernstein, 1994). Review of the item-to-total score correlations suggested that although many items correlated highly with the total score (suggesting adequate consistency), several test items had low correlations with the total score, indicating that they contributed unique variance and were not completely redundant to other items; item-to-total score correlations ranged from 0.34 to 0.84. The deletion of any single item did not result in a substantial increase in the coefficient alpha.

Test-retest reliability

Test-retest reliability is defined as the degree to which an instrument is capable of consistently measuring a phenomenon or construct over time (Anastasi and Urbina, 1997; Nunnally and Bernstein, 1994). Test-retest reliability was measured through Spearman rank-order correlations between NOS-TBI scores obtained on two consecutive days performed by the same rater (Rater 1). Due to clinical constraints and scheduling conflicts, testing on consecutive days was not always possible (mean test-retest time = 1.3 days, SD = 0.9, median = 1, range = 0–6 days). The correlation was significant ($\rho = 0.97$, $p < 0.0001$), suggesting that the NOS-TBI has a high

degree of temporal stability, albeit over a brief time interval. This time interval was selected because recovery can be rapid during inpatient rehabilitation, and this relatively rapid change would spuriously degrade the test-retest correlation coefficient if a substantially longer interval was used. A total of 49 participants were included in this analysis because one participant was unexpectedly discharged from the rehabilitation facility before the second rating could be performed (an additional reason for selecting a brief test-retest interval).

Inter-rater agreement

Inter-rater agreement is the degree to which two independent raters assign the same scores to the same observed phenomenon (Anastasi and Urbina, 1997; Nunnally and Bernstein, 1994). To reduce patient burden and ease scheduling difficulties for participants and raters, the procedure involved Rater 1 performing the assessment while Rater 2 observed. The raters did not discuss the assessment and did not comment about the participant's performance during or after the assessment to avoid biasing the other rater. In cases in which Rater 2 was uncertain of the participant's response, Rater 2 repeated the test item and scored that response. For instance, it was sometimes unwieldy for both raters to see the pupillary response, so the second rater frequently re-administered the item to adequately view the participant's response. Inter-rater agreement was measured by Cohen's kappa (range -1.0 to +1.0) for individual scale items between raters. This procedure was necessary, as the items were not expected to be equally reliable or agreeable between raters. In fact, previous studies of the NIHSS have demonstrated a considerable range of kappas (0.42–1.0) for inter-rater agreement across individual items of the scale (Brott et al., 1989; Goldstein et al., 1989; Sun et al., 2006). Further, reporting only average kappas across the scale would obfuscate such differences, leaving problematic test items undetected (von Eye and Mun, 2005). Therefore the investigators elected to calculate individual kappas, especially given that the TBI version of the scale was applied to a different neurologically-impaired population, with a high probability of resulting in unique patterns of disagreement compared to that of patients with stroke in the original NIHSS. Although the qualitative description of levels of acceptability of kappa statistics are somewhat controversial and frankly contradictory (Landis and Koch, 1977; von Eye and Mun, 2005), Fleiss (1981) recommended that kappas <0.40 should be considered "poor" agreement, kappas from 0.40–0.75 are "good" agreement, and kappas >0.75 are considered to reflect "excellent" agreement. As presented in Table 3, all of the individual item kappas were >0.75 (specifically 0.83–1.0), indicating excellent agreement.

Kendall's coefficient of concordance (KCOC) is an intraclass correlation which is capable of handling numeric ordinal data with multiple raters such as that of the NOS-TBI (Kendall, 1962). The KCOC was calculated for the NOS-TBI score between rater pairs performing ratings on the same day (von Eye and Mun, 2005). The KCOC for the NOS-TBI total score was 0.995 ($F = 203.5$, $p < 0.0001$), suggesting a very high level of inter-rater agreement for NOS-TBI total scores. It should be noted that rater pairs were highly expert, as Rater 1 (E.A.W.) trained all of the second raters; however, these raters were not necessarily representative of the non-expert raters specifically envisioned to use this scale in clinical trials of TBI

TABLE 3. UNWEIGHTED COHEN'S KAPPA STATISTICS FOR INDIVIDUAL TEST ITEMS BETWEEN INDEPENDENT RATERS

	NOS-TBI test items	κ	ASE	95% CI
1a	LOC	1.0	0.0	1.0–1.0
1b	LOC questions	0.93	0.05	0.83–1.0
1c	LOC commands	0.92	0.08	0.75–1.0
2	Gaze	0.83	0.11	0.61–1.0
3a	Visual field right	1.0	0.0	1.0–1.0
3b	Visual field left	1.0	0.0	1.0–1.0
4	Pupillary response	0.87	0.09	0.69–1.0
5a	Hearing right	0.81	0.13	0.57–1.0
5b	Hearing left	0.92	0.08	0.78–1.0
6a	Facial paresis right	0.84	0.07	0.69–0.98
6b	Facial paresis left	0.84	0.08	0.69–0.99
7a	Motor RUE	0.89	0.06	0.78–1.0
7b	Motor LUE	0.91	0.06	0.80–1.0
8a	Motor RLE	0.92	0.06	0.81–1.0
8b	Motor LLE	0.88	0.07	0.75–1.0
9a	Sensory RUE	0.86	0.10	0.67–1.0
9b	Sensory LUE	0.82	0.09	0.64–0.99
9c	Sensory RLE	0.84	0.11	0.62–1.0
9d	Sensory LLE	0.87	0.08	0.72–1.0
10	Language	0.94	0.05	0.85–1.0
11	Dysarthria	0.90	0.06	0.78–1.0
12	Neglect/extinction	0.85	0.15	0.55–1.0
13	Olfaction	0.82	0.07	0.67–0.96
14	Gait ataxia	0.88	0.06	0.75–1.0
15a	Limb ataxia right	0.96	0.04	0.89–1.0
15b	Limb ataxia left	1.0	0.0	1.0–1.0

ASE, asymptotic standard error; CI, confidence interval; LOC, level of consciousness; RUE, right upper extremity; LUE, left upper extremity; RLE, right lower extremity; LLE, left lower extremity; NOS-TBI, Neurological Outcome Scale for Traumatic Brain Injury; LOC, loss of consciousness.

interventions. Therefore, these data likely demonstrate a best-case scenario for the use of the NOS-TBI.

Convergent validity

Convergent validity is considered a subtype of construct validity in which an instrument correlates highly with other scales and constructs that are theoretically related (Anastasi and Urbina, 1997). Convergent validity was determined through Spearman rank-order correlation of the NOS-TBI score (from the initial administration by Rater 1) with the DRS, SRS, RLAS, and the FIM. The DRS, SRS, and RLAS scores were acquired during the same assessment session as the NOS-TBI; however, the FIM was administered at admission and discharge from the rehabilitation facility by trained and certified FIM administrators with no knowledge of the participant's NOS-TBI performance. Due to clinical constraints, the timing of the FIM varied from patient to patient, and occasionally only a single FIM score (admission or discharge) was available. For that reason (and for all participants), the FIM score closest to the date of the initial NOS-TBI administration was used for correlation analyses if the FIM was not administered more than ± 1 month from the NOS-TBI (mean absolute time difference = 8.4 days, SD = 7.5, median = 6.5, range 0–29 days).

Spearman correlations were calculated between the NOS-TBI score and the DRS ($\rho = 0.75$, $p < 0.0001$), SRS ($\rho = 0.59$, $p < 0.0001$), RLAS ($\rho = -0.60$, $p < 0.0001$), and the

FIM ($\rho = -0.68$, $p < 0.0001$). Only 46 participants had data available for analysis with the FIM given the ± 1 -month time limitation.

Discussion

In the present study, the convergent validity and other psychometric properties of the NOS-TBI were explored in a sample of patients undergoing inpatient rehabilitation following moderate to severe TBI. Internal consistency of the NOS-TBI was acceptable, suggesting that the NOS-TBI measures a single latent construct. This should be further assessed in future studies using factor analytic techniques, and possibly Rasch analysis (Rasch, 1966; Wright and Masters, 1982). Although evaluated over a brief time period, the test-retest reliability of the NOS-TBI was high, and this result is also an indicator of intra-rater agreement. Inter-rater reliability was thoroughly investigated. Not only was there excellent agreement between raters for the total score of the NOS-TBI, but individual item kappa statistics were uniformly in the excellent range, and exceeded those of the NIHSS (Brott et al., 1989; Goldstein et al., 1989; Sun et al., 2006). Analysis of item kappas is important, as reporting only the total score intra-class correlation statistics may conceal problematic test items that would otherwise go unnoticed. No such deficiencies were identified among the NOS-TBI test items.

Correlations between the NOS-TBI and the validation measures generally were lower than that of the NOS-TBI and the quantified neurological examination score establishing construct validity ($\rho = 0.76$; Wilde et al., 2010b), which was anticipated, as the DRS, SRS, RLAS, and FIM incorporate a number of factors in addition to neurological functioning as part of their assessment rubric. A gradient of correlation coefficients was not unexpected, as neurological dysfunction as measured by the NOS-TBI may account for varying fractions of the scores on these instruments (i.e., the SRS and RLAS are more global, unidimensional ratings of behavior than the multidimensional DRS and FIM). Additionally, the SRS was not originally designed to be used with patients in acute recovery, which may have altered the relation between it and the NOS-TBI. Also, factors other than the participant's neurological status may contribute to the level of supervision required as measured by the SRS (e.g., third-party payers, the patient's family's financial resources, and unavailability of more appropriate placement options), and this may have had an attenuating effect on the correlation coefficient. While neurological function would reasonably be expected to underlie a patient's level of functioning on the DRS, SRS, RLAS, and FIM, a substantial portion of the variance in the current sample was shared between these standard indices of functional outcome and the NOS-TBI, suggesting that the NOS-TBI is measuring a similar (but not identical) neurological construct to the DRS, SRS, RLAS, and FIM.

There are some limitations of this study that warrant brief discussion. First, it would be ideal for inter-rater agreement to be measured from separate assessments of the patient on the same day. Reducing participant burden through multiple assessments and avoiding interference in the participant's therapy schedule and family time were paramount concerns in the conduct of this study; unexpected discharges or transfers to another facility were a possibility and occurred during this study. Great care was taken so that the raters did not bias

one another by not conferring about the participant's performance, and by Rater 2 when re-administering a test item to more clearly view the participant's response. In spite of the precautions followed in this study, rater bias cannot be ruled out categorically. Future validation studies would do well to include completely independent assessment episodes. Similarly, there was potential bias in that Rater 1 performed ratings on the NOS-TBI and the validation measures. Although this is unlikely, as Rater 1 strictly followed the scoring algorithms for these measures (obtaining relevant information from the participant's chart, family members, and rehabilitation staff); however, this source of bias cannot be completely ruled out either. It should be noted that the correlation between the NOS-TBI and the FIM (when administered and scored by rehabilitation personnel with no knowledge of the NOS-TBI) was comparable to those of the other three validation measures, lessening concerns of rater bias. Future validation studies would do well to include completely independent assessment episodes and separate raters scoring the validation measures to conclusively remove this source of bias.

Second, the majority of participants in the present study were classified as having severe TBI (84%), with relatively few being classified as complicated-mild or moderate (based on Glasgow Coma Scale [GCS] score alone). It remains to be determined how useful the NOS-TBI may be in assessing neurological dysfunction in patients with mild-to-moderate TBI severity. Even if the NOS-TBI has marked floor effects in patients with mild TBI, it may yet prove useful in TBI randomized clinical trials, which frequently focus on patients with moderate-to-severe TBI.

Future studies of the NOS-TBI are needed to investigate the instrument's sensitivity to change across multiple endpoints during a patient's recovery from TBI. A similar validation technique was successfully used, in which the DRS was graded against the Glasgow Outcome Scale (GOS; Hall et al., 1985), and to determine the sensitivity to change of the Neurobehavioral Rating Scale-Revised (McCauley et al., 2001b); a similar procedure is planned for a future study by the authors of this article. Further work is required to determine how the NOS-TBI should be implemented when assigning a patient's GOS score, and more fundamentally, how NOS-TBI scores relate to GOS outcome categories. Additionally, another important property of the NOS-TBI is to determine its predictive validity compared to other measures, including post-resuscitation GCS score (Teasdale and Jennett, 1974) and post-traumatic amnesia.

In conclusion, the results of this study suggest that the NOS-TBI is a reliable and valid measure of neurological dysfunction across a broad portion of the recovery-time spectrum. Further investigation of the psychometric properties of the NOS-TBI are necessary (e.g., predictive validity and sensitivity to change), but it appears initially that this scale may be useful in clinical trials of interventions for TBI, and it may serve a complementary role to those of standard outcome measures such as the GOS and DRS.

Acknowledgments and Author Disclosure Statement

We would like to extend our gratitude to the participants and their families, whose cooperation and patience helped make this study possible. This study was supported in part

by grant NS 43353 from the National Institutes of Health/ National Institute of Neurological Disease and Stroke, to Guy L. Clifton, Principal Investigator. The information in this article and the article itself has never previously been published either electronically or in print. None of the authors have any financial or other relationships that could be construed as a conflict of interest with respect to the content of this article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- American Educational Research Association. (1999). Standards for Educational and Psychological Testing. American Psychological Association: Washington, DC.
- Anastasi, A., and Urbina, S. (1997). *Psychological Testing*, 7th ed. Prentice Hall, Inc.: Upper Saddle River, NJ.
- Boake, C. (1996). Supervision Rating Scale: A measure of functional outcome from brain injury. *Arch. Phys. Med. Rehabil.* 77, 765–772.
- Brott, T., Adams, H.P., Jr., Olinger, C.P., Marler, J.R., Barsan, W.G., Biller, J., Spilker, J., Holleran, R., Eberle, R., and Hertzberg, V. (1989). Measurements of acute cerebral infarction: A clinical examination scale. *Stroke* 20, 864–870.
- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd ed. Wiley: New York.
- Goldstein, L.B., Bertels, C., and Davis, J.N. (1989). Interrater reliability of the NIH stroke scale. *Arch. Neurol.* 46, 660–662.
- Gouvier, W.D., Blanton, P.D., LaPorte, K.K., and Nepomuceno, C. (1987). Reliability and validity of the Disability Rating Scale and the Levels of Cognitive Functioning Scale in monitoring recovery from severe head injury. *Arch. Phys. Med. Rehabil.* 68, 94–97.
- Granger, C.V., Cotter, A.C., Hamilton, B.B., Fiedler, R.C., and Hens, M.M. (1990). Functional assessment scales: a study of persons with multiple sclerosis. *Arch. Phys. Med. Rehabil.* 71, 870–875.
- Granger, C.V., Hamilton, B.B., Linacre, J.M., Heinemann, A.W., and Wright, B.D. (1993). Performance profiles of the functional independence measure. *Am. J. Phys. Med. Rehabil.* 72, 84–89.
- Hagen, C., Malkmus, D., and Durham, E. (1979). Levels of cognitive functioning, in: *Rehabilitation of the Head Injured Adult: Comprehensive Physical Management*. Professional Staff of Rancho Los Amigos Hospital: Downey, CA.
- Hall, K., Cope, D.N., and Rappaport, M. (1985). Glasgow Outcome Scale and Disability Rating Scale: comparative usefulness in following recovery in traumatic head injury. *Arch. Phys. Med. Rehabil.* 66, 35–37.
- Hall, K.M., Bushnik, T., Lakisic-Kazazic, B., Wright, J., and Cantagallo, A. (2001). Assessing traumatic brain injury outcome measures for long-term follow-up of community-based individuals. *Arch. Phys. Med. Rehabil.* 82, 367–374.
- Hamilton, B.B., Laughlin, J.A., Fiedler, R.C., and Granger, C.V. (1994). Interrater reliability of the 7-level Functional Independence Measure (FIM). *Scand. J. Rehabil. Med.* 26, 115–119.
- Heinemann, A.W., Kirk, P., Hastie, B.A., Semik, P., Hamilton, B.B., Linacre, J.M., Wright, B.D., and Granger, C. (1997). Relationships between disability measures and nursing effort during medical rehabilitation for patients with traumatic brain and spinal cord injury. *Arch. Phys. Med. Rehabil.* 78, 143–149.
- Josephson, S.A., Hills, N.K., and Johnston, S.C. (2006). NIH Stroke Scale reliability in ratings from a large sample of clinicians. *Cerebrovasc. Dis.* 22, 389–395.
- Keith, R.A., Granger, C.V., Hamilton, B.B., and Sherwin, F.S. (1987). The Functional Independence Measure: A new tool for rehabilitation. *Adv. Clin. Rehabil.* 1, 6–18.
- Kendall, M.G. (1962). *Rank Correlation Methods*, 3rd ed. Griffin: London.
- Landis, J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Linacre, J.M., Heinemann, A.W., Wright, B.D., Granger, C.V., and Hamilton, B.B. (1994). The structure and stability of the Functional Independence Measure. *Arch. Phys. Med. Rehabil.* 75, 127–132.
- McCauley, S.R., Hannay, H.J., and Swank, P.R. (2001a). Use of the Disability Rating Scale recovery curve as a predictor of psychosocial outcome following closed-head injury. *J. Int. Neuropsychol. Soc.* 7, 457–467.
- McCauley, S.R., Levin, H.S., Vanier, M., Mazaux, J.M., Boake, C., Goldfader, P.R., Rockers, D., Butters, M., Kareken, D.A., Lambert, J., and Clifton, G.L. (2001b). The neurobehavioural rating scale-revised: sensitivity and validity in closed head injury assessment. *J. Neurol. Neurosurg. Psychiatry.* 71, 643–651.
- Narayan, R.K., Michel, M.E., Ansell, B., Baethmann, A., Biegon, A., Bracken, M.B., Bullock, M.R., Choi, S.C., Clifton, G.L., Contant, C.F., Coplin, W.M., Dietrich, W.D., Ghajar, J., Grady, S.M., Grossman, R.G., Hall, E.D., Heetderks, W., Hovda, D.A., Jallo, J., Katz, R.L., Knoller, N., Kochanek, P.M., Maas, A.I., Majde, J., Marion, D.W., Marmarou, A., Marshall, L.F., McIntosh, T.K., Miller, E., Mohberg, N., Muizelaar, J.P., Pitts, L.H., Quinn, P., Riesenfeld, G., Robertson, C.S., Strauss, K.I., Teasdale, G., Temkin, N., Tuma, R., Wade, C., Walker, M.D., Weinrich, M., Whyte, J., Wilberger, J., Young, A.B., and Yurkewicz, L. (2002). Clinical trials in head injury. *J. Neurotrauma* 19, 503–557.
- Nunnally, J.C., and Bernstein, I.H. (1994). *Psychometric Theory*, 3rd ed. McGraw-Hill, Inc.: New York.
- Pallicino, P., Snyder, W., and Granger, C. (1992). The NIH stroke scale and the FIM in stroke rehabilitation. *Stroke* 23, 919.
- Powers, D.W. (2001). Assessment of the stroke patient using the NIH stroke scale. *Emerg. Med. Serv.* 30, 52–56.
- Rappaport, M., Hopkins, H.K., Hall, K., and Belleza, T. (1981). Evoked potentials and head injury. 2. Clinical applications. *Clin. Electroencephalogr.* 12, 167–176.
- Rasch, G. (1966). An item analysis which takes individual differences into account. *Br. J. Math. Stat. Psychol.* 19, 49–57.
- SAS Institute Inc. (2008). *Statistical Analysis Software for Windows*. SAS Institute, Inc.: Cary, NC.
- Schlegel, D., Kolb, S.J., Luciano, J.M., Tovar, J.M., Cucchiara, B.L., Liebeskind, D.S., and Kasner, S.E. (2003). Utility of the NIH Stroke Scale as a predictor of hospital disposition. *Stroke* 34, 134–137.
- Spilker, J., Kongable, G., Barch, C., Braimah, J., Brattina, P., Daley, S., Donnarumma, R., Rapp, K., and Sailor, S. (1997). Using the NIH Stroke Scale to assess stroke patients. The NINDS rt-PA Stroke Study Group. *J. Neurosci. Nurs.* 29, 384–392.
- Sun, T.K., Chiu, S.C., Yeh, S.H., and Chang, K.C. (2006). Assessing reliability and validity of the Chinese version of the stroke scale: Scale development. *Int. J. Nurs. Stud.* 43, 457–463.
- Teasdale, G., and Jennett, B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet* 2, 81–84.

- von Eye, A., and Mun, E.Y. (2005). *Analyzing Rater Agreement: Manifest Variable Methods*. Lawrence Erlbaum Associates: Mahwah, NJ.
- Wilde, E.A., McCauley, S.R., Kelly, T.M., Weyand, A.M., Pedroza, C., Levin, H.S., Clifton, G.L., Schnelle, K.P., Shah, M.V., and Moretti, P. (2010b). The Neurological Outcome Scale for Traumatic Brain Injury (NOS-TBI): I. Construct Validity. *J. Neurotrauma* 27, 983–989.
- Wilde, E.A., McCauley, S.R., Kelly, T.M., Levin, H.S., Pedroza, C., Clifton, G.L., Robertson, C.S., and Moretti, P. (2010a). Feasibility of the Neurological Outcome Scale for Traumatic Brain Injury (NOS-TBI) in Adults. *J. Neurotrauma* 27, 975–981.
- Williams, D.H., Levin, H.S., Eisenberg, H.M. (1990). Mild head injury classification. *Neurosurgery* 27, 422–428.
- Wright, B.D., and Masters, G.N. (1982). *Rating Scale Analysis*. MESA Press: Chicago, IL.
- Address correspondence to:
Stephen R. McCauley, Ph.D.
Cognitive Neuroscience Laboratory
Baylor College of Medicine
1709 Dryden Road, Suite 1200
Houston, TX 77030
- E-mail: mccauley@bcm.edu