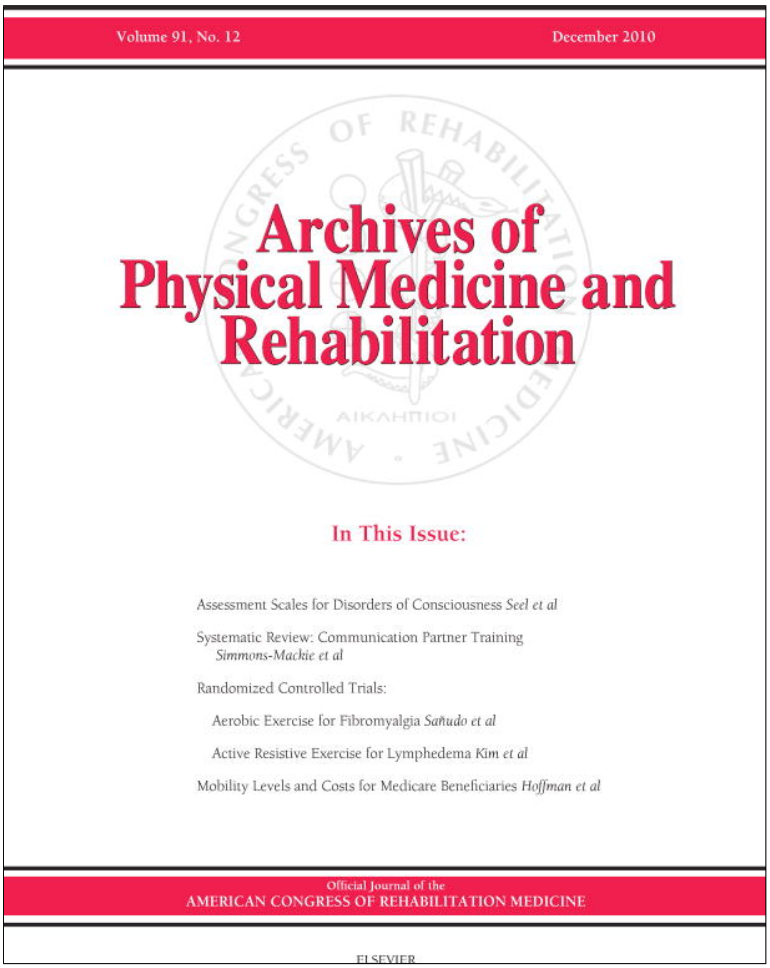

SPECIAL ARTICLE

A Practice Parameter of the American Congress of Rehabilitation Medicine

Assessment Scales for Disorders of Consciousness: Evidence-Based Recommendations for Clinical Practice and Research

Report of the American Congress of Rehabilitation Medicine, Brain Injury-Interdisciplinary Special Interest Group, Disorders of Consciousness Task Force: Ronald T. Seel, PhD, Task Force Chair, Mark Sherer, PhD, John Whyte, MD, PhD, Douglas I. Katz, MD, Joseph T. Giacino, PhD, Amy M. Rosenbaum, PhD, Flora M. Hammond, MD, Kathleen Kalmar, PhD, Theresa Louise-Bender Pape, DrPH, MA, Ross Zafonte, DO, Rosette C. Biester, PhD, Darryl Kaelin, MD, Jacob Kean, PhD, Nathan Zasler, MD

ABSTRACT. Report of the American Congress of Rehabilitation Medicine, Brain Injury-Interdisciplinary Special Interest Group, Disorders of Consciousness Task Force: Seel RT, Task Force Chair, Sherer M, Whyte J, Katz DI, Giacino JT, Rosenbaum AM, Hammond FM, Kalmar K, Pape TL, Zafonte R, Biester RC, Kaelin D, Kean J, Zasler N. Assessment scales for disorders of consciousness: evidence-based recommendations for clinical practice and research. Arch Phys Med Rehabil 2010;91:1795-1813.

Objectives: To conduct a systematic review of behavioral assessment scales for disorders of consciousness (DOC); provide evidence-based recommendations for clinical use based on their content validity, reliability, diagnostic validity, and ability to predict functional outcomes; and provide research recommendations on DOC scale development and validation.

Data Sources: Articles published through March 31, 2009, using MEDLINE, CINAHL, Psychology and Behavioral Sciences Collection, Cochrane Database of Systematic Reviews, Database of Abstracts of Reviews of Effects, Cochrane Central Register of Controlled Trials, Biomedical Reference Collection, and PsycINFO. Thirteen primary terms that defined DOC were paired with 30 secondary terms that defined aspects of measurement. Scale names, abbreviations, and authors were also used as search terms. Task force members identified additional articles by using personal knowledge and examination of references in reviewed articles.

Study Selection: Primary criteria included the following: (1) provided reliability, diagnostic validity, and/or prognostic validity data; (2) examined a cohort, case control, or case series sample of persons with DOC who were age older than or equal to 18 years; and (3) assessed in an acute care or rehabilitation setting. Articles were excluded if peer review was not conducted, original data were not reported, or an English language article was not available. The initial search yielded 580 articles. After paired rater review of study abstracts, guideline development was based on 37 articles representing 13 DOC scales.

Data Extraction: Rater pairs classified studies addressing diagnostic and prognostic validity by using the American Academy of Neurology 4-tier level of evidence scheme, and reliability by using a task force–developed 3-tier evidence scheme. An independent quality review of ratings was conducted, and corrections were made.

Data Synthesis: The Coma Recovery Scale-Revised (CRS-R), Sensory Stimulation Assessment Measure (SSAM), Wessex Head Injury Matrix (WHIM), Western Neuro Sensory Stimulation Profile (WNSSP), Sensory Modality Assessment Technique (SMART), Disorders of Consciousness Scale (DOCS), and Coma/Near-Coma Scale (CNC) have acceptable standardized administration and scoring procedures. The CRS-R has excellent content validity and is the only scale to address all Aspen Workgroup criteria. The SMART, SSAM, WHIM, and WNSSP demonstrate good content validity, containing items that could distinguish persons who are in a vegetative state, are in a minimally conscious state (MCS), or have emerged from MCS. The Full Outline of UnResponsiveness Score (FOUR), WNSSP, CRS-R, Comprehensive Levels of Consciousness Scale (CLOCS), and Innsbruck Coma Scale (INNS) showed substantial evidence of

ACRM Special Articles receive full peer review.

From the Crawford Research Institute and Brain Injury Program, Shepherd Center, Atlanta, GA (Seel, Kaelin); TIRR Memorial Hermann and Department of Physical Medicine and Rehabilitation, Baylor College of Medicine, University of Texas Medical School at Houston, Houston, TX (Sherer); Moss Rehabilitation Research Institute, Elkins Park, PA (Whyte), Department of Neurology, Boston University School of Medicine, Boston, MA (Katz); Brain Injury Program, Braintree Rehabilitation Hospital, Braintree, MA (Katz); Park Terrace Care Center, Queens, NY (Rosenbaum); Spaulding Rehabilitation Hospital, Harvard University, Boston, MA (Giacino, Zafonte); Department of Physical Medicine and Rehabilitation, Indiana University School of Medicine, Indianapolis, IN (Hammond, Kean); JFK Johnson Rehabilitation Institute, Center for Head Injuries, JFK Medical Center, Edison, NJ (Giacino, Kalmar); Research Service and the Center for Management of Complex Chronic Care Center of Excellence, Edward Hines Jr. Veterans Affairs Hospital and Department of Physical Medicine and Rehabilitation, Northwestern University Feinberg School of Medicine, Chicago, IL (Pape); Philadelphia Veterans Affairs Medical Center, Philadelphia, PA (Biester); Concussion Care Centre of Virginia and Tree of Life Services, Richmond, VA (Zasler)

Financial and technical support by the American Congress of Rehabilitation Medicine Clinical Practice Committee and the National Institute on Disability and Rehabilitation Research Model Systems Knowledge Translation Center.

No party having a direct interest in the results of the research supporting this article has or will confer a financial benefit on the authors or on any organization with which the authors are associated.

A practice parameter of the American Congress of Rehabilitation Medicine provides clinical recommendations for diagnosis, treatment, or prognosis and fully meets ACRM practice guideline development standards. The ACRM endorses all recommendations made in this practice parameter.

Giacino and Kalmar are authors of the CRS-R. Pape is author of the DOCS. Whyte was a coauthor on a study of the CRS-R. The AAN guidelines for managing panel member conflicts of interest were adopted, including the following: (a) the author panel and rater pairs were balanced with members that did not have conflicts, (b) conflicted members were not involved in the classification of studies that related to their assessment scales, and (c) conflicted members did not have responsibility for writing sections of the article that dealt with their specific assessment scales. Section lead authors were: Introduction (Giacino, Hammond); Methods (Seel, Kalmar); Question 1 Accessibility and Standardization (Rosenbaum, Kean); Question 2 Content Validity (Hammond, Kaelin); Question 3 Reliability (Sherer, Biester, Seel); Question 4 Criterion Validity (Zafonte, Katz); Question 5 Construct Validity–Rasch Model (Seel Velozo, Graves); Question 6 Diagnostic Validity (Sherer, Rosenbaum); Question 7 Prognostic Validity (Katz, Hammond); Clinical Practice Recommendations (Seel, Katz); Discussion–Practical Recommendations and Lack of Criterion Standard (Whyte, Seel); Discussion–Prognostic Validity (Pape, Seel); Research Recommendations (Kalmar, Seel, Whyte).

Correspondence to Ronald T. Seel, PhD, Shepherd Center, Crawford Research Institute, 2020 Peachtree Rd, Atlanta, GA 30309, e-mail: *ron_seel@shepherd.org*.
Reprints are not available from the author.

0003-9993/10/9112-00412$36.00/0
doi:10.1016/j.apmr.2010.07.218

internal consistency. The FOUR and the CRS-R showed substantial evidence of good interrater reliability. Evidence of diagnostic validity and prognostic validity in brain injury survivor samples had very high levels of potential bias because of methodologic issues such as lack of rater masking.

Conclusions: The CRS-R may be used to assess DOC with minor reservations, and the SMART, WNSSP, SSAM, WHIM, and DOCS may be used to assess DOC with moderate reservations. The CNC may be used to assess DOC with major reservations. The FOUR, INNS, Glasgow-Liege Coma Scale, Swedish Reaction Level Scale-1985, Loewenstein Communication Scale, and CLOCS are not recommended at this time for bedside behavioral assessment of DOC because of a lack of content validity, lack of standardization, and/or unproven reliability.

Key Words: Coma; Consciousness Disorders; Brain injuries; Diagnosis; Outcome assessment; Persistent vegetative state; Practice guidelines as topic; Prognosis; Rehabilitation; Review.

CONSCIOUSNESS CANNOT be directly observed. Therefore, clinical assessment of persons with DOC relies on observing behavior and drawing inferences about the underlying state of consciousness.[1,2] Detection of behavioral signs of consciousness is subject to interrater variability and is often confounded by unpredictable fluctuations in arousal, underlying sensorimotor impairment, unrecognized cognitive and language deficits, and sedating medications. Even when there is agreement about the behavior observed, there may be assessor variability when inferring consciousness. Diagnostic errors in classifying persons in an MCS as being in a VS have been reported to range from 30% to 40% and can have adverse consequences for clinical treatment.[3-5] Failure to detect behavioral signs of consciousness may lead to premature termination of treatment and missed clinical opportunities to establish communication, promote cognitive and functional progress, and identify and manage pain. Conversely, misinterpreting nonpurposeful or reflexive behavior as conscious behavior may lead to falsely optimistic prognoses, excessively prolonged or aggressive treatment, and delays in adequately planning for long-term disability. In the most severe circumstances, misdiagnosis can cause inappropriate family and legal decisions regarding withdrawal of life-sustaining treatment. Although neuroimaging and electrophysiologic procedures are evolving as potential components of the DOC clinical assessment, they do not have sufficient evidentiary support to be included in formal diagnostic criteria or routine clinical care.[6-11]

List of Abbreviations

AAN	American Academy of Neurology
ACRM	American Congress of Rehabilitation Medicine
BI-ISIG	Brain Injury-Interdisciplinary Special Interest Group
CI	confidence interval
CLOCS	Comprehensive Levels of Consciousness Scale
CNC	Coma/Near-Coma Scale
CPC	Clinical Practice Committee
CRS-R	Coma Recovery Scale-Revised
DOC	disorders of consciousness
DOCS	Disorders of Consciousness Scale
DRS	Disability Rating Scale
FOUR	Full Outline of UnResponsiveness Score
GCS	Glasgow Coma Scale
GLS	Glasgow-Liege Coma Scale
IC	internal consistency
ICC	intraclass correlation coefficient
INNS	Innsbruck Coma Scale
IRR	interrater reliability
LOEW	Loewenstein Communication Scale
MCS	minimally conscious state
MSKTC	Model Systems Knowledge Translation Center
NIDRR	National Institute on Disability and Rehabilitation Research
PVS	persistent vegetative state
RLS85	Swedish Reaction Level Scale-1985
SMART	Sensory Modality Assessment Technique
SSAM	Sensory Stimulation Assessment Measure
TRR	test–retest reliability
VS	vegetative state
WHIM	Wessex Head Injury Matrix
WNSSP	Western Neuro Sensory Stimulation Profile

To provide a common frame of reference for the diagnosis and treatment of persons with DOC, consensus-based diagnostic guidelines have been developed to define 3 ascending levels of disordered consciousness: coma,[12] the VS,[13] and the MCS.[14] Coma represents a state of unarousable unresponsiveness in which there is no evidence of self-awareness or environmental awareness.[12] The eyes remain continuously closed, purposeful responses to environmental stimuli cannot be elicited, and there is no evidence of discrete localizing responses or language comprehension and expression. The defining characteristic of coma is the absence of spontaneous eye opening or sleep/wake cycles. Behavior is limited to reflexive activity, indicating failure of both the reticular activating system and integrated cortical activity.[15]

In persons with persistent unconsciousness, spontaneous or stimulus-induced eye opening often reemerges. The recovery of eye opening with continued absence of observable signs of cognitively mediated behavior signals a transition to VS. The AAN Multi-Society Task Force on PVS outlines 3 diagnostic criteria that must all be met to establish the diagnosis of VS:

1. No evidence of sustained, reproducible, purposeful or voluntary behavioral responses to visual auditory, tactile, or noxious stimuli.
2. No evidence of language comprehension or expression.
3. Intermittent wakefulness manifested by the presence of sleep/wake cycles (ie, periodic eye opening).[16]

In persons with VS, autonomic functions are usually sufficiently preserved to support cardiorespiratory functions and permit survival. There is usually gradual resumption of spontaneous or elicited movement; however, this is always nonpurposeful or reflexive. Vocalizations or facial expressions such as smiling and crying may occur in VS but not in the context of meaningful environmental interaction.[16-18]

Unlike those in coma and VS, persons in MCS demonstrate minimal but definitive behavioral evidence of self-awareness or environmental awareness. Conscious behaviors are often subtle, occur inconsistently, and must be carefully differentiated from reflexive or random behavior. The Aspen Workgroup first proposed and subsequently refined MCS diagnostic criteria with a case definition published in 2002.[14,19] To establish the diagnosis of MCS, there must be clear and reproducible evidence of 1 or more of the following behaviors:

1. Simple command following.
2. Gestural or verbal yes/no responses (regardless of accuracy).
3. Intelligible verbalization.

4. Movements or affective behaviors that occur in contingent relation to relevant environmental stimuli and are not attributable to reflexive activity. Exemplars include the following:
   A. Episodes of crying, smiling, or laughter in response to the linguistic or visual content of emotional but not neutral topics or stimuli.
   B. Vocalizations or gestures that occur in direct response to the linguistic content of comments or questions.
   C. Reaching for objects that demonstrates a clear relationship between object location and direction of reach.
   D. Touching or holding objects in a manner that accommodates the size and shape of the object.
   E. Visual pursuit or sustained fixation that occurs in direct response to moving or salient stimuli.

Aspen Workgroup criteria for emergence from MCS require reliable demonstration of either interactive communication or functional object use.[14] Reliable interactive communication requires accurate yes/no responses to at least 6 situational orientation questions (eg, "Are you sitting in a chair?") and may occur through verbalization, gesture, or assistive technology. To demonstrate reliable functional object use, appropriate use of at least 2 different items must be observed on at least 2 different occasions (eg, comb brought to the head and toothbrush to the mouth).

Although consensus-based diagnostic guidelines for coma, VS, and MCS have been widely accepted in the United States and internationally, there are no corresponding guidelines to inform the DOC assessment approach. Behavioral assessment scales for DOC are frequently used for diagnosis, prognosis, and treatment planning. To address the content and prognostic limitations of the ground-breaking GCS,[20-23] second-generation DOC scales have proliferated since the early 1990s. However, these second-generation DOC scales demonstrate marked procedural variability, ranging from brief unstructured bedside scales to standardized assessment strategies with explicit rating criteria. Recent evidence suggests that standardized scales may be more accurate than unstructured approaches in detecting consciousness when the assessment protocol is yoked to consensus-based diagnostic criteria and relies on operationally defined administration and scoring procedures.[5] To date, there has been no systematic evaluation of the standardization, content, psychometric properties, and clinical utility of second-generation DOC behavioral assessment scales.

To address this critical need, the ACRM BI-ISIG DOC Task Force conducted a systematic, evidence-based literature review to evaluate the extent that each DOC scale (1) has content that differentiates VS, MCS, and emergence from MCS; (2) produces reliable ratings between examiners and over time; (3) generates valid diagnostic findings; and (4) predicts functional outcomes. Recommendations are made to guide clinical practice and research on the development and validation of DOC scales.

METHODS

DOC Task Force Expert Panel

The task force is made up of 14 members from the ACRM BI-ISIG who have expertise in brain injury and DOC. Eight members have significant expertise in scale development and validation. Six are past or current members of the ACRM CPC, which performs quality standards reviews for ACRM-generated evidence-based guidelines.

Oversight and Funding

The project plan was developed in accordance with the AAN Process Manual for the Development of Evidence-Based Guidelines[24] and was preapproved by the ACRM CPC and the Archives of Physical Medicine and Rehabilitation editorial staff. The ACRM CPC and the NIDRR MSKTC provided financial and technical support.

Literature Search

The literature search identified articles published through March 31, 2009, by using EBSCO and selecting the following databases: MEDLINE, CINAHL, Psychology and Behavioral Sciences Collection, Cochrane Database of Systematic Reviews, Database of Abstracts of Reviews of Effects, Cochrane Central Register of Controlled Trials, and Biomedical Reference Collection. In addition, PsycINFO was searched. Thirteen primary search terms were used to define DOC: akinetic mutism, apallic syndrome, coma AND vegetative state, coma AND post-head injury, coma AND post-trauma, coma AND post-traumatic, coma AND traumatic, minimally conscious, minimally responsive, persistent vegetative state, post-head injury coma, prolonged post-traumatic unawareness, and unawareness state. Each of the 13 primary terms was paired with 30 secondary terms that defined aspects of measurement: assessment, classification, course, diagnosis, diagnostic, evaluation, injury severity score, instrument, measure, natural history, neurologic examination, observer variation, outcome, predictive, prognosis, prognostic, progression, psychometric, questionnaire, recovery, reliability, reproducibility of results, scale, sensitivity, specificity, test, tool, trauma severity indices, validation, and validity. Filter terms (eg, plant, animal, religion, ethics) were used to eliminate irrelevant articles. An additional search was conducted by using scale names, abbreviations, and author names as search terms. Finally, task force members used personal knowledge of DOC scale articles and examined references in reviewed articles to identify additional relevant articles.

Selection of Articles and Scales

Selected articles met the following criteria: (1) provided reliability, diagnostic validity (including criterion validity), and/or prognostic validity data on a second-generation DOC behavioral assessment scale; (2) examined a cohort, case control, or case series sample made up of persons with DOC resulting from traumatic brain injury, stroke, and/or other nontraumatic brain injury etiology with most age greater than or equal to 18 years; and (3) assessed the sample in either an acute care or a rehabilitation setting. Articles were excluded if peer review was not conducted, original data were not reported, or an English language article was not available. Prognostic validity studies were excluded if the only outcome predicted was survival versus death. To assure that selected DOC scales are in active use, scale inclusion in the systematic review required at least 1 peer-reviewed article published between 1996 and 2009 and at least 1 peer-reviewed article that reported either diagnostic or prognostic validity data. If a revised DOC scale had articles that met selection criteria, then the earlier version was excluded.

Overall, the literature search yielded 580 articles of interest for which the lead author and research coordinator prescreened abstracts to exclude articles that did not possess data. Of these 580 articles, 494 were excluded. The remaining 86 abstracts, representing 22 scales, were formally screened for inclusion by task force member pairs. Forty-four were excluded during the abstract screen,[25-68] and an additional 4 were excluded during

Table 1: Description of Disorders of Consciousness Scales

Scale	Behavioral Content	No. of Scales (No. of Items)*	Item Response Set†	Score Range and Interpretative Guidelines‡	Estimated Time Required (min)
CNC	Visual, auditory, command following, threat response, olfactory, tactile, pain, vocalization	8 (11)	"Occurs 2–3 times," "occurs 1–2 times," or "does not occur"	Average item score: 0.00–0.89=no coma; 0.90–2.00=near coma; 2.01–2.89=moderate coma; 2.90–3.49=marked coma; 3.50–4.00=extreme coma	10
CRS-R	Auditory, visual, motor, oral, communication, arousal	6 (23)	"Absent" or "present"	Total score=0–23	25
CLOCS	Eye responses, motor, posture, communication, general responsiveness	7 (7)§	Varies per item, 5–9 anchored responses	Total score=0–42	5
DOCS	Auditory, visual, tactile, sensory, swallowing, olfactory	1 (23)	"None," "generalized," or "localized"	Logit transformed total score=0–100	45
FOUR	Eye response, motor response, respiration, brainstem reflexes	4 (4)	5 Anchored responses	Total score=0–16	10
GLS	Eye, verbal, motor, brainstem reflexes	4 (4)	Varies per item, 4–6 anchored responses	Total score=3–20	5
INNS	Eye responses, auditory, pain, posture, oral	1 (8)	Varies per item, 3–4 anchored responses	Total score=0–23	10
LOEW	Mobility, respiration, visual, auditory, communication	5 (25)	5 Anchored responses	Total score=0–100	25
RLS85	Responsiveness	1 (1)	8 Levels of "reaction"	1–3=responsive; 4–8=not responsive	15
SMART	Auditory, vision, tactile, olfactory, gustatory, wakefulness, motor, communication	8 (8)	5 Anchored responses	Each scale score=1–5 MCS or higher if rated a 5 on a sensory modality on 5 consecutive administrations	60+
SSAM	Auditory, vision, tactile, olfactory, gustatory, eye opening, motor, vocalization	5 (15)	6 Anchored responses	Total score=15–90	30
WHIM	Basic behaviors, social/communication, attention/cognitive, orientation/memory	4 (58)	"Absent" or "present"	Total score=0–58	30–120‖
WNSSP	Visual, tactile, olfactory, arousal/attention, auditory, expressive communication	5 (32)	Varies per item, 3–6 anchored responses	Total score=0–110	45

*No. of items refers to scored items; multiple stimuli may be presented before a single item is rated.
†Most scales use anchored (specifically defined) response sets (rating choices) that are individualized to each item and range from lack of behavior to specific volitional behavior.
‡Higher total scores indicate higher levels of consciousness for all scales except CNC and RLS85, in which higher scores indicate lower levels of consciousness.
§CLOCS originally contained 8 single-item scales, but authors recommend deleting scale 2.
‖Up to 58 items ordered hierarchically from easier to harder may be administered. Item administration is discontinued after 10 consecutive incorrect responses, which can lead to large variations in administration times.

the full article review.[22,69-71] Thus, guideline development was based on 37 articles[5,21,23,72-106] with the following 13 DOC scales meeting inclusion criteria: CRS-R, CNC, CLOCS, DOCS, FOUR, GLS, INNS, LOEW, RLS85, SMART, SSAM, WHIM, and WNSSP. See table 1 for a brief description of each scale. Authors of 12 scales used classical test theory methodologies to provide evidence of reliability and validity. The authors of the DOCS used the Rasch model to provide evidence of reliability and validity.

Rating the Evidence

Several evidence rating schemes were used. Scale standardization was rated as acceptable or unacceptable based on DOC task force expert consensus between paired reviewers that procedures were sufficiently defined to facilitate consistent administration and scoring of items. DOC task force paired reviewers also provided consensus ratings on content validity as excellent, good, acceptable, or unacceptable based on the extent that each scale's items covered the Aspen Workgroup's 4 criteria for transitioning from VS to MCS and 2 criteria for emerging from MCS.

Three aspects of reliability were examined: IC, IRR, and TRR. No validated scheme was identified for rating medical assessment scales' reliability. A subcommittee of DOC task force members in consultation with a biostatistician synthesized recommendations and findings from the reliability research literature and created a 3-tier reliability classification system.[107-125] Reliability evidence was abstracted, and each

Table 2: DOC Scales: Standardization of Procedures, Interpretive Guidelines, and Evaluation of Item Content Based on Aspen Workgroup Criteria

Scale	Author (y)	Public Domain	Std Admin/Scoring	Dx Guide	Aspen WC Diff Dx	Aspen WC VS→MCS	Aspen WC MCS→Emerg
CRS-R	Giacino (2004)	Y	Y	Y	Y	4	2
SSAM	Rader (1994)	Y	Y	N	N	4	1
WNSSP	Ansell (1989)	Y	Y	N	N	4	1
SMART	Gill-Thwaites (1999)	N	Y	Y	N	4	1
WHIM	Shiel (2000)	N	Y	N	N	4	1
DOCS	Pape (2005)	Y	Y	N	N	3	0
CNC	Rappaport (1992)	Y	Y	Y	N	3	0
CLOCS	Stanczak (1984)	Y	N	N	N	3	0
LOEW	Borer-Alafi (2002)	Y	N	N	N	3	0
RLS85	Stalhammar (1988)	Y	N	Y	N	3	0
FOUR	Wijdicks (2005)	Y	N	N	N	2	0
GLS	Born (1985)	Y	N	N	N	2	0
INNS	Benzer (1991)	Y	N	N	N	1	0

Abbreviations: Aspen WC Diff Dx, whether the scale provides guidelines for using items or scale scores to make differential diagnoses among VS, MCS, and emerged based on Aspen Workgroup criteria; Aspen WC MCS→Emerg, the number of Aspen Workgroup Criteria (of 2 possible) for differential diagnosis of MCS from emerged that is addressed by each scale's item content; Aspen WC VS→MCS, the number of Aspen Workgroup criteria (of 4 possible) for differential diagnosis of VS from MCS that is addressed by each scale's item content; Dx Guide, diagnostic guidelines provided to interpret scale scores beyond higher or lower total scores, indicating higher or lower levels of consciousness; N, no; Public Domain, whether scale can be accessed for free or is available by purchase only; Std Admin/Scoring, whether adequate procedures are provided for standardized administration and scoring; Y, yes.

study was rated class I (low risk of bias), II/III (moderate to high risk of bias), or IV (very high risk of bias) based on methodologic features (appendices 1-3). IC, IRR, and TRR coefficients were rated excellent, good, acceptable, or unacceptable (appendices 1b and 1c). Reliability conclusions were based on a synthesis of the evidence, with the strength of each conclusion rated as established as reliable, probably reliable, or possibly reliable based on AAN guideline development nomenclature.

Criterion, diagnostic, and prognostic validity evidence was abstracted from studies and rated class I (low risk of bias), II (moderate risk of bias), III (high risk of bias), or IV (very high risk of bias) by using the AAN classification of evidence scheme (appendix 4).[24] Two non–task force members (see Acknowledgements) served as expert reviewers for 1 article that used the Rasch model to validate the DOCS and rated the evidence class I, II, III, or IV. Criterion validity correlation coefficients were rated weak (.30–.49), moderate (.50–.74) or strong (≥.75). For diagnostic and prognostic validity studies, if sensitivity, specificity, and 95% CIs were not reported in an article but sufficient raw data were available, then reviewers calculated these statistics. Validity conclusions were based on a synthesis of the evidence with the strength of each conclusion rated established as valid, probably has validity, or possibly has validity based on AAN guideline development nomenclature.

By using structured forms, reviewer pairs independently rated articles, discussed disagreements, and reached consensus. In a few cases, disagreements were referred to a third reviewer who served as a tiebreaker. The NIDRR-funded MSKTC reviewed the accuracy of the task force ratings on the 16 articles that generated the most evidence. Discrepancies between MSKTC and task force ratings were researched, and corrections were made. Ratings and data from the finalized article reviews were then entered into evidence tables, which were used for synthesizing the evidence, forming conclusions, and generating recommendations.

Recommendations for clinical use of DOC scales were based on evidence of standardization, content validity, reliability, and criterion/construct validity by using a modified AAN strength of recommendation scheme as follows: may be used with minor reservations, may be used with moderate reservations, may be used with major reservations, not recommended at this time, or not recommended. The ACRM governing board approved the final version of the article.

RESULTS

Question 1: Which DOC Scales Are Accessible, Provide Standardized Administration and Scoring Procedures, and Provide Interpretive Guidelines?

Evidence. Eleven of the 13 scales can be accessed for free, whereas the WHIM and SMART must be purchased (table 2). Prerequisite to purchasing the SMART is submission of a work-based portfolio and completion of a 5-day training course held in the United Kingdom. Nine of the 13 scales can be administered in 30 minutes or less. The DOCS and the WNSSP take approximately 45 minutes, whereas the WHIM and SMART can take 1 hour or more to complete, depending on severity of impairment and the number of items or stimuli administered.

Experts' consensus-based evaluations concluded that the CRS-R, SSAM, WHIM, WNSSP, SMART, DOCS, and CNC have well-defined administration and scoring procedures that facilitate consistent use. The administration and scoring procedures of the CLOCS, LOEW, RLS85, FOUR, INNS, and GLS were considered insufficiently standardized to be consistently applied. Only the CRS-R, SMART, CNC, and RLS85 provided interpretive guidelines for scores. The CRS-R provided diagnostic guidelines based on the Aspen Workgroup criteria for VS, MCS, and emerged from MCS. The SMART defines scoring criteria that differentiate MCS or higher from VS; individual ratings on each scale are translated into Rancho scale levels. The CNC translates an average item score into diagnostic categories (no coma, near coma, moderate coma, marked coma, or extreme coma) that are not consistent with Aspen Workgroup consensus-based diagnostic classifications. Similarly, the RLS85 provides interpretive guidelines (alert, drowsy, unconscious) that do not fit Aspen Workgroup diagnostic classifications.

Conclusions. The CRS-R, CNC, CLOCS, RLS85, SSAM, WHIM, WNSSP, DOCS, FOUR, GLS, INNS, and LOEW are accessible and can be administered in a reasonable time. The

SMART may in practice be inaccessible for non–United Kingdom practitioners because of prerequisite training requirements and costs. The CRS-R, SSAM, WHIM, WNSSP, SMART, DOCS, and CNC have acceptable standardized administration and scoring procedures. The CLOCS, LOEW, RLS85, FOUR, INNS, and GLS have unacceptable standardization of administration and scoring procedures. Only the CRS-R met all criteria for accessibility, standardization, and interpretive guidelines that fit Aspen Workgroup consensus-based diagnostic classifications.

Question 2: To What Extent Does Each Scale's Item Content Address Key Distinguishing Features of DOC as Defined by the Aspen Workgroup?

Evidence. Content validity provides essential evidence that a DOC scale's items cover a representative sample of DOC-relevant behavior. The Aspen Workgroup criteria for transitioning from VS to MCS and for emerging from MCS were used as the operational definition of the DOC construct. Experts' consensus-based evaluations concluded that the item content of the CRS-R, SSAM, WHIM, WNSSP, and SMART assesses all 4 VS versus MCS criteria (see table 2). The DOCS, CNC, CLOCS, LOEW, and RLS85 have items that cover 3 of the 4 MCS diagnostic criteria. The FOUR, GLS, and INNS cover only 1 or 2 of the 4 MCS diagnostic criteria.

As for assessing emergence from MCS, the CRS-R contains items that cover the 2 Aspen Workgroup emergence criteria. The SMART, SSAM, WHIM, and WNSSP have items that cover 1 of the 2 emergence criteria. The CLOCS, CNC, DOCS, FOUR, GLS, INNS, LOEW, and RLS85 did not contain items that address emergence from MCS criteria.

Conclusions. The CRS-R has excellent content validity and is the only scale to address all Aspen Workgroup criteria. The SMART, SSAM, WHIM, and WNSSP demonstrate good content validity, containing items that could differentiate persons who are VS, MCS, or emerged from MCS. The DOCS, CNC, CLOCS, LOEW, and RLS85 have acceptable content validity but only for differentiating persons in MCS from those in VS based on Aspen Workgroup criteria. The FOUR, GLS, and INNS have unacceptable content validity because their items do not sufficiently cover a representative sample of DOC behavior.

Question 3: Which DOC Scales Reliably Measure Behavior?

Evidence—IC. IC measures the extent that scale items are interrelated and produce similar ratings. IC does not provide evidence that DOC scale items represent a unidimensional construct. The Cronbach alpha is typically used to establish IC, and coefficients greater than or equal to .80 are typically indicative of good IC. The Cronbach alpha is somewhat dependent on the number of scale items such that, if a 4-item scale with a Cronbach alpha of .75 were increased to 8 items while maintaining the exact level of item interrelatedness, the Cronbach alpha would increase to .90 purely because of the increase in items.[107,108,116,122] The Cronbach alpha is not meaningful for scales with fewer than 4 items.[123,125]

Six class I and 2 class II/III studies provided IC evidence on 7 DOC scales (table 3). Two class I studies (N=80 and 120) of the 4-item FOUR found Cronbach alphas of .95 and .86, respectively.[80,81] One class I study (N=57) of the 32-item WNSSP showed a Cronbach alpha of .95.[21] One class I study (N=101) of the 7-item CLOCS found a Cronbach alpha of .88, whereas 1 class I study (N=80) of the 23-item CRS-R showed the Cronbach alpha to be .83.[72,75] One class I study (N=84) of the 8-item INNS found the Cronbach alpha to be .78.[88] One class II/III study of the 23-item DOCS showed a Cronbach alpha of .85 in persons (N=68) with DOC resulting from closed head injury.[77] One class II/III study of the CNC with limited representativeness (N=20) indicated Cronbach alphas less than or equal to .65 when administered at 1, 8, and 16 weeks postinjury.[23] No IC studies were found for the GLS, LOEW, SSAM and WHIM. IC is not an appropriate gauge of reliability for the SMART and RLS85 because these scales generate single-item scores for clinical decision making.

Conclusions—IC. The FOUR has established excellent IC (multiple class I). The WNSSP probably has excellent IC (class I). The CRS-R and CLOCS probably have good IC (class I). The INNS probably has acceptable IC (class I). The DOCS possibly has good IC (class II/III). The IC of the CNC is possibly unacceptable (class II/III). IC is unproven for the GLS, LOEW, SSAM, and WHIM (not studied).

Evidence—IRR. IRR refers to the degree of agreement between 2 or more raters when using a DOC scale. Several research designs can be used to produce quality IRR evidence,

Table 3: Design Characteristics and Outcomes in Studies With IC Analyses for DOC Scales

Scale	Author (y)	Evidence Class	IC Rating	N Size	N Rep	No. of Items	Statistic Calculated*	Results
FOUR	Wolf (2007)	I	Excellent†	80	Y	4	CA	.95
FOUR	Wijdicks (2005)	I	Excellent†	120	Y	4	CA	.86
WNSSP	Ansell (1989)	I	Excellent	57	Y	32	CA	.95
CLOCS	Stanczak (1984)	I	Good	101	Y	7	CA	.88
CRS-R	Giacino (2004)	I	Good	80	Y	24	CA	.83
INNS	Diringer (1997)	I	Acceptable	84	Y	8	CA	.78
DOCS	Pape (2005)	II/III	Good	68 (CHI)	Y	23	CA	.85
				27 (OBI)	Y	23	CA	.77
CNC	Rappaport (1992)	II/III	Unacceptable	20	N	11	CA	.43 (wk 1)
							CA	.65 (wk 8)
							CA	.65 (wk 16)

Abbreviations: CA, Cronbach α; CHI, closed head injury sample; Evidence Class, Task Force classification system for rating risk of bias in the internal consistency methodology: I=low risk of bias, II/III=moderate to high risk of bias, and IV=very high risk of bias; IC rating, strength of IC α coefficient: unacceptable $<.70$, acceptable=.70–.79, good=.80–.89, excellent$\geq$.90; N, no; N Rep, whether the sample was representative enough to determine internal consistency; N Size, sample size; OBI, other brain injury sample; Y, yes.

*All CAs were calculated by using all items on each scale.

†For scales with 4 to 6 items, CA coefficients of $\geq$.80 are considered excellent.

Table 4: Design Characteristics and Outcomes in Studies With IRR Analyses for DOC Scales

Scale	Author (y)	Evidence Class	Reliability Rating	N Size	N Suff	No. of Admin	Time Btw Admin	No. of Raters*	Stats Calc	Results
FOUR	Wolf (2007)	I	Good	80	Y	2	1h	2	ICC	.96
									κ_w	.85
FOUR	Wijdicks (2005)	I	Good	120	Y	2	1h	2	κ_w	.82
FOUR	Stead (2009)	II/III†	Excellent	69	Y	2	10min	2	ICC	.98
									κ_w	.89
CRS-R	Giacino (2004)	II/III	Good	20	N	2	Same day	2	Sr	.84
									Wilcoxon	P=.10
CRS-R	Schnakers (2008)	II/III	Good	24	N	1	0	2	κ	.80
SMART	Gill-Thwaites (2004)	II/III†	Excellent	60	Y	1	0	2	ICC	.96
LOEW	Borer-Alafi (2002)	II/III	Excellent	22	N	1	0	2	κ	.90
RLS85	Stalhammar (1988)	II/III	Unacceptable	81	N	2	<25min	2	κ	.69
RLS85	Tesseris (1991)	II/III	Acceptable	74	N	2	<20min	2	κ	.73
GLS	Born (1987)	II/III	Unacceptable	30	N	1	0	2	κ	.69 Reflexes
										.65 Motor
WHIM	Shiel (2000)	IV	Good	25	N	1	0	2	κ_{mean}	.86
WHIM	Majerus (2000)	IV	Good	5	N	1	0	2	κ_{mean}	.84
CNC	Rappaport (1992)	IV	Excellent (Sys error not eval)	20	N	1	0	2	Sr	.98 (wk 1)
										.98 (wk 8)
										.97 (wk 16)
CLOCS	Stanczak (1984)	IV	Excellent (Sys error not eval)	20	N	1	0	3	Pr	.96 (median score)
WNSSP	Ansell (1989)	IV	Excellent (Sys error not eval)	23	N	1	0	3	Pr	.94–.99
SSAM	Rader (1989; 1994)	IV	Excellent (Sys error not eval)	19	N	1	0	2	Pr	.89

Abbreviations: Evidence Class, Task Force classification system for rating risk of bias in IRR methodology: I=low risk of bias, II/III=moderate to high risk of bias, IV=very high risk of bias; κ_{mean}, mean of κ's; κ_w, weighted κ; N, no; No. of Admin, whether single (1) or separate (2) administrations of the test were conducted; No. of Raters, the number of raters compared; N Size, sample size; N Suff, whether sample size was sufficient to produce a CI with a width of .20 for the calculated reliability coefficient; Pr, Pearson correlation; Reliability Rating, strength of the calculated IRR coefficient: unacceptable<.70, acceptable=.70–.79, good=.80–.89, excellent≥.90; Sr, Spearman correlation; Sys error not eval, systematic error not evaluated; Time Btw Admin, the time between scale administrations; Y, yes.
*All raters were reported to be blinded except Shiel (2000), who provided conflicting information on rater blinding.
†Samples were narrow or skewed.

including a single administration of scale items, with at least 2 observers providing independent ratings of each patient's responses or 2 independent administrations and ratings of scale items within a very short time frame (eg, minutes, at most hours). Interpreting IRR coefficients is not straightforward, and low coefficients may not be purely attributable to insufficient standardization of the scale's administration and scoring procedures. Potential nonscale factors such as raters' duration of DOC clinical experience and level of training and experience when using the scale, true patient variability when 2 independent administrations and ratings of scale items occur, and the sample's range on the factor can all be implicated in low IRR coefficients.

Two class I, 8 class II/III, and 7 class IV studies provided IRR evidence for 12 scales (table 4). Two class I studies of the FOUR in which separate independent administrations and ratings of items were conducted within 1 hour found average weighted kappa coefficients of .82 and .85[80,81] and an average ICC of .96.[81] One class II/III study of the FOUR with a sample that contained many persons who were fully conscious produced higher levels of agreement than the class I studies (ICC=.98; weighted κ=.89).[82] Two small sample class II/III studies of the CRS-R by using different methodologies and statistical approaches provided consistent evidence of IRR (Spearman r=.84, no systematic error, eg, no differences in magnitude of ratings between pairs; κ=.80).[72,73]

One class II/III study of the SMART in a predominantly VS sample that used a single administration of scale items with 2 independent raters produced an ICC of .96.[100] One small sample class II/III study of the LOEW by using a single administration of scale items with 2 independent raters reported a κ of .90.[90] Two class II/III studies of the RLS85 by using the same methodology reported IRRs just above and below the cutoff score for acceptability (κ=.69 and .73).[93,95] One small sample class II/III study of the GLS by using a single administration of scale items with 2 independent raters found low levels of agreement on reflex ratings (κ=.69) and motor ratings (κ=.65).[86]

Small sample class IV studies of the CNC, CLOCS, WNSSP, and SSAM all reported excellent correlation values (range, .89–.99), indicating little random error in ratings.[21,23,75,101,102] However, these class IV studies did not evaluate systematic error, introducing an unacceptably high risk of bias. Class IV studies generated inconclusive findings for the WHIM and DOCS because of failure to either implement or report appropriate IRR methodology.[77,103,104] No IRR study was found for the INNS.

Conclusions—IRR. The FOUR has established good IRR (multiple class I). The CRS-R probably has good IRR (multiple class II/III). The SMART and the LOEW possibly have excellent IRR (class II/III). The GLS possibly has unacceptable IRR (class II/III). The IRR of the RLS85 is unproven (inconsistent class II/III). IRR is unproven for the WHIM, CNC, CLOCS, WNSSP, SSAM, and DOCS (class IV) and INNS (not studied).

Evidence—TRR. TRR assumes that the construct measured does not change over time and calculates the level of within-subject agreement on repeated scale administrations. Selecting a time interval in which repeated DOC scale admin-

Table 5: Design Characteristics and Outcomes in Studies With Criterion Validity Analyses for DOC Scales

Scale*	Author (y)	Evidence Class	Criterion Rating	Sample Size (Spec)	Ref Std Ind	Raters Masked	Stat Calc	Ref Std1	Coeff1 (95% CI)	Ref Std2	Coeff2 (95% CI)
CLOCS	Stanczak (1984)	III	Strong	101 (W)	N	I	Pr	Nursing Scale†	.75 (.65 to .82)	—	—
CLOCS	Johnston (1996)	IV	Moderate to strong	43 (W)	Y	NS	Pr	GCS	.90 (.87 to .93)	Neuro dx scale‡	−.49 (−.61 to −.39)
RLS85		III	Strong	74 (W)	N	M	Sr	GCS-R1	−.88 (−.92 to −.82)	E2CS-R1	.92 (.88 to .95)
RLS85	Tesseris (1991)		Strong				Sr	GCS-R2	−.76 (−.84 to −.64)	E2CS-R2	.90 (.85 to .94)
RLS85	Starmark (1988)	IV	Strong	47 (W)	N	N	Sr	GCS-R1	−.94 (−.97 to −.89)	GCS-R2	−.96 (−.98 to −.93)
RLS85	Matousek (1996)	IV	Moderate	34 (NS)	Y	I	*r*	EEG§	.57 (.29 to .76)	EEG§ (CV: age) EEG§ (CV: meds)	.59 (.31 to .77) .50 (.19 to .72)
CRS-R	Giacino (2004)	IV	Strong	80 (W)	N (CRS) Y (DRS)	N	Sr	CRS	.97 (.95 to .98)	DRS	−.90 (−.93 to −.85)
CRS-R	Schnakers (2008)	IV	Moderate to strong	77 (W)	Y	N	Sr	WHIM	.76 (.65 to .84)	FOUR	.63 (.47 to .75)
CNC	Rappaport (1992)	IV	Moderate	20 (W)	Y	N	Sr	DRS	.69 (.36 to .87)	MEPA	.52 (.12 to .85)
CNC	Talbot (1994)	IV	Strong	7 (Nar)	Y	N	Sr	DRS	.94 (.66 to .99)	—	—
SSAM	Rader (1989; 1994)	IV	Moderate	20 (Nar)	Y	NS	Pr	GCS	.70 (.37 to .87)	DRS	−.61 (−.83 to −.23)
SMART	Gill-Thwaites (2004)	IV	Weak to moderate	60 (Nar)	Y	N	Pr	WNSSP	.70 (.54 to .81)	RLAS	.47 (.25 to .65)
WHIM	Majerus (2000)	IV	Strong	23 (Nar)	Y	NS	Sr	GCS (BL)	.83 (.64 to .93)	GCS (Final)	.95 (.88 to .98)
WNSSP	Ansell (1989)	IV	Moderate	57 (W)	Y	N	Kr	RLAS	.73 (.58 to .83)	—	—

Abbreviations: —, no other analyses; (BL), baseline assessment; Coeff (95% CI), coefficient reported in study with the 95% CI generated from reviewers' statistical calculation denoted in parentheses; Criterion Rating, magnitude of the correlation coefficients: weak=.30–.49, moderate=.50–.74, strong≥.75; Evidence Class, AAN classification of the risk of bias in study results: I=low risk of bias, II=moderate risk of bias, III=moderate to high risk of bias, IV=very high risk of bias; E2CS, Edinburgh-2 Coma Scale; Ind, Independent; meds, administered sedating medications 0.5 to 6.0 hours prior to EEG; MEPA, Multisensory Evoked Potential Abnormalities; Kr, Kendall rank correlation; N, no; Neuro dx Scale, Neuroradiologic Ordinal Brain Damage Rating; NS, not stated/could not be determined; N Size, sample size; Pr, Pearson correlation; Raters Masked, whether the researcher administering the reference standard was blind to the investigational scale score: M=masked, I=independent raters but either not masked or masking could not be determined, N=not masked; *r*, partial correlation; R1, rater 1; R2, rater 2; Ref Std, reference standard used as the comparison DOC measure; (Spec), the spectrum of persons with DOC represented in the sample based on the reference standard scores: wide (W)=broad representation of DOC, narrow (N)=limited range of DOC, Sr, Spearman rank correlation; Stat Calc, statistic calculated; Y, yes.

*All criterion validity analyses compared total scores on the scale being evaluated with the reference standard score.

†Internal standardized 7-point scale of nurses' ratings of patient consciousness.

‡Author rating of blind neuroradiologic and/or encephalographic severity on a 9-point scale.

§EEG frequency converted to a 12-point scale; CV, defined covariate in the analysis.

istrations would be expected to produce equivalent scale scores is critical. For example, when persons in VS or MSC are early in their recovery course and receiving treatment, improvement is likely, and the interval between scale administrations would need to be no more than a few days.

Two class II/III and 3 class IV studies provided TRR evidence for 5 scales. One class II/III study of the CRS-R by using a small but representative sample 22 to 169 days postinjury found high levels of agreement (Spearman r=.94) and no systematic error in ratings (Wilcoxon test, P=.10) when independent scale administrations were performed 36 hours apart.[72] One class II/III study of the SMART by using a predominantly VS sample 27 to 3120 days postinjury in which scale administrations were repeated 1 day apart produced an ICC of .97.[100] Single class IV studies of the CLOCS and SSAM reported good to excellent correlation values but did not evaluate the potential for systematic error.[75,101] One class IV study of the WHIM generated acceptable κ levels, but ambiguities with regard to independence of ratings and how much time elapsed between scale administrations introduced a high risk of bias.[103] No TRR studies were found for the CNC, DOCS, FOUR, GLS, INNS, LOEW, RLS85, or WNSSP.

Conclusions—TRR. The CRS-R and the SMART possibly have excellent TRR (class II/III). TRR is unproven for the WHIM, CLOCS, and SSAM (class IV) and the CNC, DOCS, FOUR, GLS, INNS, LOEW, RLS85, and WNSSP (not studied).

Question 4: To What Extent Do DOC Scales Demonstrate Criterion Validity?

Evidence. Criterion validity refers to the degree that behavior on an investigational DOC scale corresponds to behavior on an established DOC measure, which is called a reference standard. Two class III studies and 11 class IV studies provided criterion validity evidence for 8 scales (table 5). All 13 studies measured behavior on the investigational DOC scale and a reference standard concurrently. One class III study of the CLOCS in which independent raters used a standardized 7-point scale as a reference standard found a Pearson correlation of .75 (95% CI, .65–.82).[75] One class III study of the RLS85 by using a masked rating procedure with the GCS as a reference standard found high Spearman correlations between 2 rater pairs (absolute 95% CI ranging between .82–.92 and .64–.84).[95]

Nine of the 11 class IV studies reported very high correlations (absolute coefficient value ≥.69) between the investigational scale and a reference standard. Research designs that lacked independent or masked raters introduced a very high risk of bias, which may have contributed to the high correlations reported in these studies. For example, in 7 class IV studies involving the CRS-R,[72,73] CNC,[23,74] WNSSP,[21] RLS85,[94] and SMART,[100] the same rater completed the investigational scale and a reference standard. In 3 class IV studies involving the CLOCS,[76] SSAM,[101,102] and WHIM,[104] masking

status was not stated. Other research design issues in class IV studies included small samples,[23,73,101,104] samples with narrow DOC representation,[92,100] and use of reference standards with item content that highly overlapped (>70%) the investigational scales.[72,75,94,95] No criterion validity studies were found for the DOCS, FOUR, GLS, INNS, and LOEW.

Conclusions. The CLOCS and the RLS85 possibly have strong criterion validity (class III). Criterion validity is unproven for the CRS-R, CNC, SSAM, SMART, WHIM, and WNSSP (class IV) and for the DOCS, FOUR, GLS, INNS, and LOEW (not studied).

Question 5: What Scales Demonstrate Construct Validity Based on the Rasch Model?

Evidence. DOCS and WHIM items are ordered in a difficulty hierarchy representing a single dimension, presumably of consciousness. Persons who successfully complete more difficult items are assumed to have higher levels of consciousness. The Rasch model analyzes the extent to which items conform to a unidimensional difficulty hierarchy. Factor analytic techniques establish evidence of unidimensionality, and fit statistics calibrate and evaluate how well item difficulty and person ability conform to those of a linear interval scale measure. The Rasch model sets a higher standard for achieving construct validity compared with classical test theory's preferred methods for establishing criterion validity.

One class III study[77] of the DOCS found some evidence of a unidimensional difficulty hierarchy when using the Rasch model. The rating scale model was adequate. The results of the principal components analysis indicated that unexplained variance in the item residuals was negligible. After 11 of 34 items were removed, outfit statistics for the closed head injury sample indicated good item fit on 21 of 23 items. Person separation of 2.38 for the closed head injury sample suggested that 3 reliable strata existed. There was also evidence that the DOCS did not fit the Rasch model. Lack of item invariance was an issue, with 11 of 34 items removed from the scale because of differential item functioning over time. Further evidence of lack of item invariance was the reordering of items based on injury type ("closed" vs "other" head injury). Outfit statistics for the "other" head injury sample indicated overfit in 14 of 23 items.

A number of methodologic decisions limited the DOCS study's ability to provide evidence of unidimensionality and model fit. Infit statistics are considered most indicative of model fit; however, none were reported. Infit statistics are also typically used to decide which items require removal because of lack of model fit; however, differential item functioning results were used to remove items that were unstable over time. Last, separate scales were formed for "closed" versus "other" brain injury because the 2 subsamples had mean scores on different ends of the brain injury severity dimension; when using the Rasch model, severity is not a theoretic basis for forming separate scales.

No studies were identified that used the Rasch model to examine the unidimensional hierarchic structure of the WHIM.

Conclusions. The DOCS possibly has unidimensional hierarchic interval characteristics when used with persons with closed head injury (class III). The unidimensional hierarchic structure of the DOCS in other acquired brain injury subgroups has not been adequately established (class IV). The unidimensional hierarchic structure of the WHIM is unproven (not studied).

Question 6: How Well Do DOC Scales Differentiate Diagnostic Levels of Persons With DOC?

Evidence. Diagnostic validity refers to a DOC scale's ability to establish an accurate diagnosis compared with the true diagnosis, which is measured by a reference standard. For example, an investigational DOC scale is used to identify persons who are in a VS versus an MCS, and these diagnostic results are compared with the true diagnoses, which could be established through use of another DOC scale or consensus agreement between 2 or more clinicians relying on bedside observations. Sensitivity and specificity values are typically calculated to provide better controlled indices of diagnostic accuracy than the percentage of correct diagnoses made, which can be affected by base rates that may be idiosyncratic to a given sample.

Three class IV studies provided diagnostic validity evidence for 2 scales. For all 3 studies, task force reviewers had to calculate sensitivity and specificity coefficients by using raw data provided in the study. One class IV study[91] of the RLS85 in which the same rater used a reference standard (GCS) with item content that highly overlapped the RLS85 had high levels of sensitivity and specificity (≥.90). Two class IV studies of the CRS-R[72,73] both used unmasked raters and had high levels of sensitivity (≥.97) but less than optimal specificity (.66–.76) when using the DRS and the WHIM as reference standards to compare VS and MCS diagnoses. No diagnostic validity studies were found for the CNC, CLOCS, SSAM, SMART, WHIM, WNSSP, DOCS, FOUR, GLS, INNS, and LOEW.

The lack of a criterion standard measure for DOC complicates the interpretation of diagnostic validity findings (see Discussion). Additional data collection and analyses are recommended to determine the validity of investigational DOC scale versus reference standard diagnoses (see Recommendations for Research #10).

Conclusions. Diagnostic validity is unproven for the CRS-R and RLS85 (class IV) and the CNC, CLOCS, RLS85, SSAM, SMART, WHIM, WNSSP, DOCS, FOUR, GLS, INNS, and LOEW (not studied).

Question 7: How Well Do DOC Scales Predict Outcomes?

Evidence. Prognostic validity establishes the degree that DOC scale scores or diagnoses demonstrate utility (eg, explain significant variance) in predicting survivors' recovery of consciousness or function. To establish a prognostic model of survivor outcomes, ideally, multiple variables, including the DOC scale, would be studied to identify a parsimonious set of predictors that maximally explain variance.

DOC prognostic studies varied widely with regard to the timing of predictor scale assessment relative to the onset of brain injury, the outcome measures used, and the timing of the outcome assessment (table 6). For example, the timing of the predictor DOC scale assessment varied from the day of acute admission in 7 studies,[81,82,84,85,87,88,91] to the acute care period in 4 studies,[75,76,80,83] to subacute and chronic periods up to 5 years postinjury in 4 studies.[77-79,90] With regard to outcome measures, 6 studies used the Glasgow Outcome Scale,[75,76,84,85,87,91] 4 studies dichotomized Rankin Scale scores,[80-83] and other studies used dichotomized recovery of consciousness[77] and rehabilitation readiness[90] or discharge settings.[88] The timing of outcome measurement greatly varied

Table 6: Methodologic Issues and Statistical Outcomes in Studies With Prognostic Validity Analyses for DOC Scales

Scale	Author (y)	Evidence Class	Method Issues	Predictor Score	Timing of Pred Meas	Ref Std Outcome	Timing of FUP Meas	Stats Calc Coefficient	(95% CI)	OddsR (95% CI)	% Correct	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)
FOUR	Wolf (2007)	I		Total	≤24h	Rankin Good Rankin Mod-Death	30d	LR	—	.58 (.41–.82)	—	NS	NS	NS	NS
FOUR	Stead (2009)	III	N	Total	≤24h	Rankin Good Rankin Mod-Death	Acute DC	LR	—	.43† (.26–.71)	—	NS	NS	NS	NS
FOUR	Wijdicks (2005)	IV	M	Total ≥9	ICU stay	Rankin Good Rankin Mod-Death	3mo	LR	—	.84 (.77–.92)	—	.75 (NS)	.76 (NS)	NS	NS
FOUR	Eken (2009)	IV	M	Total	ICU stay	Rankin Good Rankin Mod-Death	3mo	AUC	.75† (.68–.81)	—	—	NS	NS	NS	NS
INNS	Diringer (1997)	I		Total (minus Oral Automism Score)	≤24h	Death Nursing home Home w/asst Independent	3mo (post-DC)	WL	.72†	—	74% 0% 0% 56%	NS	NS	NS	NS
GLS	Born (1985)	III	I	Total + age Total + age	≤24h	GOS Good-Mod GOS SEV-VS GOS Good-Mod GOS SEV-Death	6mo 6mo	LR LR	— —	NS† NS†	— —	.00 (.00–.28) .92 (.80–.97)	1.00 (.96–1.00) .79 (.65–.88)	.00 .80 (67–.89)	.88 (.80–.93) .92 (.79–.97)
GLS	Born (1988)	III	I	Motor + reflex	≤24h	GOS Good-Mod GOS SEV-VS GOS Death	6mo	LR	—	NS*	83% 0% 86%	NS	NS	NS	NS
GLS	Mukherjee (2000)	IV	M	Total	≤24h	GOS Good-Mod GOS SEV-VS GOS Death	Acute DC	MR	NS	—	84% (all groups)	NS	NS	NS	NS
DOCS	Pape (2005)	IV	M, F	Total ≥48 Total <48	≤94d	Conscious Not conscious	1y	LR	—	1.27 (.97–1.66)	—	.71 (.54–.84)	.71 (.44–.89)	.84 (.66–.94)	.52 (.31–.73)
DOCS	Pape (2006)	IV	M, F	DOCS average, first evaluation, slope	21–365d	CHART scales	1y	LR	—	NS	—	—	—	—	—
DOCS	Pape (2009)	IV	M, F	DOCS change ≤4 DOCS change >4	Second minus first assessment	Conscious Not conscious	4mo 8mo 1y	AUC AUC AUC	.91 (NS) .88 (NS) .80 (NS)	—	—	.84 (NS) .83 (NS) .69 (NS)	.87 (NS) .83 (NS) .70 (NS)	.94 (NS) .91 (NS) .69 (NS)	.68 (NS) .70 (NS) .70 (NS)
LOEW	Borer-Alafi (2002)	IV	M, F	Total ≥50 Total <50	26–632d	Rehab ready Not rehab ready	26–632d	LR	—	1.12† (1.04–1.20)	—	.81 (.63–.92)	.90 (.54–.99)	.96 (.79–1.00)	.60 (.33–.83)
RLS85	Johnstone (1993)	IV	M	Level 1–3 Level 4–8	Acute admission	GOS Good GOS Mod-VS	Acute DC	—	—	—	—	.96 (.91–.98)	.28 (.17–.42)	.81 (.75–.86)	.68 (.45–.85)
CLOCS	Stanczak (1984)	IV	M, F, N	Level	<3d	Modif. GOS (9 levels)	Acute DC	MR	.32 (NS)	—	—	—	—	—	—
CLOCS	Johnston (1996)	IV	M, F, N	Level (log-transformed)	<3d	Modif. GOS (9 levels)	Acute DC	MR	.07† (NS)	—	—	—	—	—	—

Abbreviations: —, no other analyses; % Correct, the proportion of patients correctly classified by the predictor groups (only reported if sensitivity etc. could not be calculated); CHART, Craig Handicap Assessment and Reporting Technique; DC, discharge; Evidence Class, AAN classification of the risk of bias in study results: I=low risk of bias, II=moderate risk of bias, III=moderate to high risk of bias, IV=very high risk of bias; Method Issues, F=loss to follow-up ≥20%, I=independent raters for predictor and reference standard but no masking or masking not stated, M= lack of masking-same person administered reference standard and predictor scale or masking not stated, N=narrow sample composition or not stated; FUP, follow-up; Good-Mod, good recovery or moderately disabled; GOS, Glasgow Outcome Scale; modif. GOS, modified Glasgow Outcome Scale; ICU, intensive care unit; NPV, negative predictive value (% true negative); NS, not stated; OddsR, odds ratio; PPV, positive predictive value (% true positive); Rankin, modified Rankin Scale; Ref Std Outcome, the reference standard and classification used to determine outcome groups; Rehab, rehabilitation; Sens, sensitivity calculation; Sev-VS, severely disabled or vegetative state; Spec, specificity calculation; Stats Calc, type of statistic calculated: AUC=area under the curve ratio, LR=logistic regression, Mod-Death, moderate disability, moderately severe disability, severe disability, or death; MR=multiple regression, WL=Wilkes λ; Timing of FUP Meas, timing of the outcome measurement relative to the injury date unless noted; Timing of Pred Meas, timing of predictor measurement relative to the injury date unless noted; GOS, Glasgow Outcome Scale; Good, no symptoms, no significant disability or slight disability.

*$P<.05$ of odds ratio or coefficient.

†$P<.01$.

across studies from time of acute care discharge to 4 years after onset. Thus, evidence of a DOC scale's prognostic validity must be interpreted within the parameters of the predictor and outcome assessment time frames and the target outcome.

Two class I, 2 class III, and 13 class IV studies provided prognostic validity evidence for 9 scales. Most studies predicted disability levels with death included as an outcome. Few studies examined the prognosis of survivors on functional outcomes of primary interest (eg, emerged from DOC or rehabilitation ready at 3 or 6 months postinjury; functional independence or supervision/care needs at 6 or 12 months postinjury). One class I study of the FOUR[81] administered within 24 hours of injury found that the total score differentiated persons with good recovery from persons who were disabled or died at 30 days postinjury (odds ratio .58 [95% CI, .41–.82]). One class III study of the FOUR[82] administered within 24 hours of injury to a sample made up of many persons who were conscious also found that the total score differentiated persons with good recovery from persons who were disabled or died at acute care discharge (odds ratio .43; [95% CI, .26–.71]). One class I study of the INNS[88] administered within 24 hours of injury found that the total score demonstrated modest accuracy for predicting death versus independent home placement but no ability to classify accurately persons discharged home with assistance or to a nursing home at 3 months postacute care discharge. Two class III studies of the GLS[85,87] reported mixed evidence with predictive utility dependent upon inclusion of death as an outcome. For example, the GLS administered within 24 hours of injury demonstrated strong positive predictive (.80 [95% CI, .67–.89]) and negative predictive values (.92 [95% CI, .79–.97]) when differentiating between persons with good recovery or moderate disability and persons who were severely disabled, in a vegetative state, or deceased.[87] However, when persons who died were removed from the analysis, the positive predictive value was .00 when differentiating persons with good recovery or moderate disability from persons who were either severely disabled or in a VS.[87]

Commonly encountered research design issues in the 13 class IV studies included lack of masking between predictor and outcome ratings in 12 studies,[75-80,83,84,90,91,96,105] greater than 20% loss to follow-up in 6 studies,[75-79,90] and a narrow sample composition in 3 studies.[75,76,82] Two class IV studies of the SMART and WNSSP (not included in table 6 because of space restrictions) used concurrent scores from the investigational scale as the predictor and outcome measures.[96,105] No prognostic validity studies were found for the CRS-R, CNC, CLOCS, SSAM, and WHIM.

Conclusions. The FOUR administered 24 hours or sooner postinjury is probably predictive of good recovery versus disability or death at 30 days postinjury (class I). The INNS administered 24 hours or sooner postinjury is probably not predictive of independent living versus disability at 3 months after acute care discharge (class I). The GLS administered 24 hours or sooner postinjury is possibly not predictive of good recovery or moderate disability versus severe disability or PVS at 6 months postinjury (class III). The GLS administered 24 hours or less postinjury is possibly predictive of good recovery or moderate disability versus severe disability or PVS or death at 6 months postinjury (class III). Prognostic validity is unproven for the RLS85, SMART, WNSSP, DOCS, and LOEW (class IV) and the CRS-R, CNC, CLOCS, SSAM, and WHIM (not studied).

Recommendations for Clinical Practice

Table 7 summarizes the evidence and strength of conclusions for the standardization, content validity, reliability, criterion/construct validity, diagnostic validity, and prognostic validity of DOC behavioral assessment scales. Only class IV evidence exists to support DOC scales' abilities to make discrete diagnoses or to predict restoration of consciousness or functional outcome in brain injury survivors. Thus, recommendations for or against the clinical use of each DOC scale to make diagnostic classifications or provide postacute prognoses for brain injury survivors must be deferred until class I to III evidence becomes available. The following recommendations for clinical use of DOC scales reflect a synthesis of best evidence, with a focus on standardization, content validity, reliability, and criterion/construct validity.

1. The CRS-R *may be used to assess DOC with minor reservations*. This recommendation is supported by expert consensus that the CRS-R has excellent content validity and acceptable standardized administration and scoring procedures. Studies provide evidence that the CRS-R probably has good IRR and good IC and possibly has excellent TRR. Criterion validity for the CRS-R is unproven.
2. The SMART, WNSSP, SSAM, WHIM, and DOCS *may be used to assess DOC with moderate reservations*. This recommendation is supported by expert consensus that the SMART, WNSSP, SSAM, WHIM, and DOCS have either good or acceptable content validity and acceptable standardized administration and scoring procedures. Each scale has limited evidence with regard to reliability or criterion validity. One study provides evidence that the SMART possibly has excellent interrater and TRR. However, criterion validity for SMART subscale scores is unproven, and significant purchase and training costs may be prohibitive for clinicians outside the United Kingdom. One study provides evidence that the WNSPP probably has excellent IC, but IRR, TRR, and criterion validity are unproven. The WHIM and the SSAM lack evidence of IRR, IC, TRR, and criterion validity. The unidimensional hierarchic item structure and scoring system of the WHIM have not been studied, and the scale must be purchased. One study provides evidence that the DOCS for persons with closed head injury possibly has good IC and possibly has construct validity based on the Rasch model; however, interrater and TRR are unproven.
3. The CNC *may be used to assess DOC with major reservations*. This recommendation is supported by expert consensus that the CNC has acceptable content validity and acceptable administration and scoring procedures. However, 1 study provides evidence that the IC of the CNC is possibly unacceptable; IRR, TRR, and criterion validity are unproven.
4. The RLS85, LOEW, and CLOCS are *not recommended at this time* for serial bedside behavioral assessment of DOC. Although expert consensus indicates that these scales have acceptable content validity, administration and scoring procedures are not sufficiently standardized and probable evidence does not exist for their reliability or criterion validity.
5. The FOUR, INNS, and GLS are *not recommended* for serial bedside behavioral assessment of DOC. Expert consensus indicates that these scales' items do not sufficiently cover a representative sample of DOC behavior and have unacceptably low levels of standardization in their administration and scoring procedures.

Table 7: Summary of Evidence and Strength of Conclusions for DOC Behavioral Assessment Scales

Scale	Q1 Standardized Admin/Scoring	Q2 Content Validity	Q3a IC	Q3b IRR	Q3c TRR	Q4 & Q5 Criterion Validity	Q6 Diagnostic Validity	Q7 Prognostic Validity
CRS-R	Acceptable	Excellent	Good (class I)	Good (multiple class II/III)	Excellent (class II/III)	Unproven (class IV)	Unproven (class IV)	Unproven (not studied)
SMART	Acceptable	Good	NA	Excellent (class II/III)	Excellent (class II/III)	Unproven (class IV)	Unproven (not studied)	Unproven (class IV)
WNSSP	Acceptable	Good	Excellent (class I)	Unproven (class IV)	Unproven (class IV)	Unproven (class IV)	Unproven (not studied)	Unproven (class IV)
SSAM	Acceptable	Good	Unproven (not studied)	Unproven (class IV)	Unproven (class IV)	Unproven (class IV)	Unproven (not studied)	Unproven (not studied)
WHIM	Acceptable	Good	Unproven (not studied)	Unproven (class IV)	Unproven (class IV)	Unproven (class IV)	Unproven (not studied)	Unproven (not studied)
DOCS	Acceptable	Acceptable	Good (class II/III)	Unproven (class IV)	Unproven (not studied)	Construct valid* (class III)	Unproven (not studied)	Unproven (class IV)
CNC	Acceptable	Acceptable	Unacceptable (class II/III)	Unproven (class IV)	Unproven (not studied)	Unproven (class IV)	Unproven (class IV)	Unproven (class IV)
CLOCS	Unacceptable	Acceptable	Good (class I)	Unproven (class IV)	Unproven (class IV)	Strong (class III)	Unproven (not studied)	Unproven (class IV)
LOEW	Unacceptable	Acceptable	Unproven (not studied)	Excellent (class II/III)	Unproven (not studied)	Unproven (not studied)	Unproven (not studied)	Unproven (class IV)
RLS85	Unacceptable	Acceptable	NA	Unproven (class IV)	Unproven (not studied)	Strong (class III)	Unproven (class IV)	Unproven (class IV)
FOUR	Unacceptable	Unacceptable	Excellent (multiple class I)	Good (multiple class I)	Unproven (not studied)	Unproven (not studied)	Unproven (not studied)	Probably predictive (class I) 30d postinjury Good vs Mod-Death
INNS	Unacceptable	Unacceptable	Acceptable (class I)	Unproven (not studied)	Unproven (not studied)	Unproven (not studied)	Unproven (not studied)	Probably not predictive (class I) 3mo postdischarge Independent vs disability
GLS	Unacceptable	Unacceptable	Unproven (not studied)	Unacceptable (class II/III)	Unproven (not studied)	Unproven (not studied)	Unproven (not studied)	Possibly not predictive (class III) 6mo postinjury Good-Mod vs Sev VS Possibly predictive (class III) 6mo postinjury Good-Mod vs Sev-Death

Abbreviations: multiple class I, established as reliable, diagnostic valid, or prognostic valid; class I study or multiple class II or II/III studies, probably reliable, diagnostic valid, or prognostic valid; class II, II/III, or III studies, possibly reliable, diagnostic valid, or prognostic valid; class IV or not studied, unproven as reliable, diagnostically valid, or prognostically valid. Admin, administration (and scoring procedures); NA, not applicable.
*Construct valid for unidimensional hierarchical structure for closed head injury population.

DISCUSSION

Practical Recommendations When Assessing With DOC Scales

In addition to each scale's psychometric properties, other factors can influence the accuracy of DOC scale use. Inconsistent purposeful behavior is a core feature of MCS; thus, extended or repeated assessment with a DOC scale is likely to improve diagnostic accuracy. Clinicians should have training and experience with the DOC population to facilitate diagnostic accuracy when using a DOC scale. For example, sensory and physical deficits can confound the diagnosis of consciousness. Although scale administration procedures may attempt to address confounds to varying degrees, an experienced assessor familiar with the clinical variations exhibited in the DOC population may be more likely to identify and navigate confounding influences than a novice armed only with scale administration instructions. Receiving formal training on a DOC scale's administration and scoring procedures before clinical use would seem to be a prerequisite for facilitating diagnostic accuracy. Almost all DOC scale developers recommend training before clinical use. When formal training is offered, it varies widely, from watching videotapes to attending a week-long training course. No studies have systematically assessed whether prior DOC scale training or expertise promotes improved assessment accuracy or whether the amount of training needed to use a DOC scale differs based on experience or discipline.

Lack of a Criterion Standard Measure to Establish Diagnostic Validity

There is no validated objective or subjective criterion standard DOC measure on which to establish a true diagnosis, which complicates the interpretation of diagnostic validity findings. Consider a CRS-R study[72] in which the DRS was the reference standard used to establish the true diagnosis. CRS-R–generated MCS diagnoses had a high level of sensitivity but less than optimal specificity compared with the DRS diagnoses. Does this low level of specificity indicate that the CRS-R generates a high number of false-positive MCS diagnoses, or does the DRS fail to identify higher levels of conscious functioning because of less specific DOC content and a shorter scale administration window? The finding that any DOC scale identifies more cases of MCS than VS compared with another assessment method by itself does not provide sufficient evidence of better diagnostic sensitivity.

Use of consensus-based expert diagnoses of patients as in either a VS or an MCS might not resolve the criterion standard issue given that clinicians using unstructured observations consistently diagnose 30% to 40% of MCS patients as vegetative. Because the performance of persons with DOC can vary within a single day, behavioral evidence of their consciousness level will depend, in part, on how frequently, how long, and how carefully the clinician observes as well as the care taken to rule out or accommodate confounding sensory or motor deficits. Trained, experienced clinicians using structured DOC behavioral assessment scales likely outperform expert diagnosticians, but the evidence will have to come from convergent, independent sources (see Recommendations for Research #10).

Baseline and Outcome Measurement in Prognostic Validity Research

Ideally, DOC scales as primary predictor variables should be measured across patients at a similar time point. Yet homogeneous timing of the DOC scale predictor assessment can be logistically challenging because persons early in their recovery may be residing in a variety of settings, including acute care, a specialty DOC rehabilitation program, a skilled nursing facility, or at home. Random variation in the performance of persons with DOC is typical; therefore, using a single DOC scale assessment as the predictor may prove less accurate than using multiple assessments to optimize prognostic value. DOC scale total scores relative to time postinjury, rates of change in total scores, threshold score values, and discrete diagnoses all warrant study for optimal prognostic value. DOC scale scores or diagnoses would not be expected solely to predict recovery but should demonstrate utility (eg, explain significant variance) in predicting recovery. Ideally, multiple potential predictors (eg, DOC scale scores, age, an anoxic component, injury location) would be studied to identify a parsimonious set that maximally explains outcome variance.

Minimizing measurement error in outcomes can be challenging in the DOC population, particularly when the outcome measured is status of consciousness. Telephone assessment of DOC is difficult, and there is no validated interview that differentiates VS from MCS. Further, it is difficult to expect families to observe and report reliably and validly their loved one's status of consciousness when clinicians have a 30% to 40% error rate. Conducting an assessment at the house of the person with DOC or bringing the person into the clinic for an assessment is often logistically or economically infeasible. The duration of follow-up for a DOC prognostic study must be sufficient to detect differences. Yet, as the duration between baseline and follow-up measurement increases, study costs and loss to follow-up may also increase.

Providing accurate and meaningful prognostic information about persons with DOC to families and clinicians early in the recovery course is a critical concern. Prognostic research requires a DOC sample of sufficient size to study adequately a large number of predictor variables and have generalizability. Extramural funding for multicenter research designs is almost a necessity for DOC prognostic research. Functional outcome measurement requires balancing the risk of floor and ceiling effects for a population that can have widely disparate long-term outcomes. Self-care, mobility, communication, and participation outcomes are meaningful to all persons surviving severe brain injury and can be targeted in addition to global indices of disability.[126-131]

Study Limitations

A few limitations should be noted. We did not use the search engine EMBASE to identify articles or include articles that had no English language translation. Our review did not examine the ability of DOC scales to detect meaningful change. The review is current with studies published through March 31 2009. As new evidence emerges and conceptualizations of DOC change, revisions to this systematic review and practice guideline will be required.

Recommendations for Research

Based on our systematic review, the following recommendations address critical issues that would improve DOC scale development and validation research. Implementation of some recommendations would benefit from extramural funding and multicenter research designs.

1. Our literature review indicated that few studies articulated a specific operational definition or theoretic framework for the construct disordered consciousness that served as the basis for creating DOC scale item content. Scales and items should reflect testable a priori definitions of a DOC

construct and cover the full range of behaviors that differentiate current consensus-based understanding of DOC levels. Finally, scales and items should undergo periodic revisions to keep up to date with evidence.

2. Studies should use patient samples whose consciousness levels are diverse enough to cover the range of behaviors that DOC scales are intended to measure. Samples also should be of sufficient size to generate findings with narrow CIs.
3. Because of proximity and/or medical condition, many persons cannot travel without great expense (eg, thousands of dollars) for prognostic follow-up evaluations. Research is needed to examine the relative cost and accuracy of several remote assessment options, including examination by local health professionals with expert technical support, family assessment with telephone training and support, or remote patient assessment with a video link.
4. Authors should provide a detailed description of the scale administration methods used and include raters' levels of clinical DOC experience as well as DOC scale–specific training received and level of experience.
5. Use of masked raters on the investigational scale and reference standard assessment is critical to manage the risk of inflated agreement in scores, discrete diagnoses, and prognoses. In criterion and diagnostic validity studies, randomizing the administration order of 2 scales while using the same rater does not resolve the masking issue. Ideally, trained raters other than the scale developers or principal investigators should be used to establish external validity.
6. When using expert consensus on a clinical diagnosis as the reference standard, the methods should detail the number and duration of experts' assessments, the extent that their examinations covered Aspen Workgroup and/or other criteria, and the steps taken to control for potential sensory or motor confounds. The method should also detail how disagreements in independent assessments were resolved.
7. When DOC scale items are ordered on an assumed continuous dimension, the Rasch model should be used to provide evidence that the scale is unidimensional, items conform to an interval hierarchy, and changes in scores are equivalent along the continuum (eg, a 5-point change on the low and high ends of the scale constitute the same amount of change).
8. Statistical analyses for IRR and TRR studies should take into account the item rating scheme used, sample normality assumptions, and assess for both random and systematic error.
9. To test the validity of diagnoses or prognoses, 2-by-2 contingency tables are created in which sensitivity, specificity, positive and negative predictive values, and likelihood ratios can be calculated. The percentage of correct diagnoses or prognoses is subject to sample bias, and reporting only this percentage does not provide sufficient evidence to establish validity.
10. To interpret false-positive (low specificity) and false-negative (low sensitivity) results in diagnostic validity analyses, the discrepant cases should be further compared with an independent, indirect standard outcome such as 1-month postbaseline functional level, 1-month postbaseline rate of cognitive improvement, or concurrent physiologic studies. If the discrepant cases diagnosed as a false-positive MCS when using the investigational DOC scale have outcomes that converge with the outcomes of the congruently diagnosed MCS cases, then diagnostic validity would be established.
11. DOC scale scores are sometimes used to detect meaningful change along a continuum or to establish discrete diagnostic classifications indicative of meaningful transitions. Our review did not specifically examine the ability of DOC scales to detect meaningful change. Each use has a different validation approach. To detect change along a continuum, analyses must define meaningful change in relation to the standard error of measurement, clinical meaningfulness, and prognosis. To examine diagnoses made based on scale cutoff scores, criterion reference reliability analyses should be conducted.

Acknowledgments: We thank Rachel Emery, BA; Stacie Waddell, MLIS, Mary Kaye Howard, MEd, and Christine Willis, MLIS, and; Laura Vazquez, MPH (Shepherd Center); Craig A. Velozo, PhD, OTR/L, and Daniel E. Graves, PhD, FACRM (Statistical Consultants and Reviewers of Rasch Model Study); Patricia Brown, EdD, Kurt L. Johnson, PhD, Mark Harniss, PhD, Karen Cook, PhD, Katie Snappin, MS, Grace Wang, MS, Katherine Schomer, MA, (MSKTC Quality Review); Marcel P. J. M. Dijkers, PhD, FACRM, and Mark V. Johnston, PhD (ACRM CPC Review); and Gary S. Gronseth, MD, FAAN (Consultant on AAN Systematic Review Process).

APPENDIX 1: DOC TASK FORCE EVIDENCE CLASSIFICATION AND RATING SCHEME OF RELIABILITY

Class	IC Methodologic Features Associated With Each Evidence Class
I	Low risk of bias. A study design in which the DOC sample is representative, internal consistency analyses are conducted on the scale or subscales used to make clinical decisions and the appropriate statistic (typically Cronbach α) is calculated.
II/III	Moderate or high risk of bias. A study design in which the DOC sample is narrow (eg, 90% are VS) OR a less stringent internal consistency statistic is used.
IV	Very high risk of bias. Any design in which the DOC sample is made up of subjects within a single consciousness level (eg, all are VS) OR the sample is skewed with nonrepresentative subjects (eg, 75% emerged from MCS) OR IC statistics are calculated for the total score when subscale scores are used to make clinical decisions.

Class	IRR and TRR Methodologic Features Associated With Each Evidence Class
I	Low risk of bias. A study design with a representative DOC sample of sufficient size to produce results within a 0.2 CI band. An appropriate scale administration method is used. For IRR: 1 scale administration and $\geq$2 independent observations OR video of scale administration and $\geq$2 independent observations OR 2 independent administrations and 2 independent observations within the same day. For TRR, an appropriate methodology to answer the test–retest question, notably an appropriate time between ratings. Appropriate statistics are used: (a) for nominal/dichotomous data, Cohen κ; (b) for normally distributed, interval data, ICCs; and (c) for skewed/nonnormally distributed data, Spearman rank correlation plus analysis of systematic differences between raters (eg, Bland-Altman plot, Wilcoxon signed rank test).

APPENDIX 1: DOC TASK FORCE EVIDENCE CLASSIFICATION AND RATING SCHEME OF RELIABILITY (Cont'd)

Class	IRR and TRR Methodologic Features Associated With Each Evidence Class
II/III	Moderate or high risk of bias. A study design in which a narrow sample of the DOC population is used OR the sample size is not sufficient to produce results within a 0.2 CI band. Method used to administer and score tests introduces a moderate to high degree of bias. Statistical approach introduces a moderate to high degree of bias.
IV	Very high risk of bias. Any design in which the sample is made up of a single consciousness level (eg, all vegetative state) OR the sample is highly skewed with nonrepresentative subjects (eg, 75% emerged from MCS). Method used to administer and score tests introduces a very high degree of bias such as a lack of independence in observation ratings. Statistical approach introduces a very high degree of bias (eg, using correlation statistics without analyzing systematic differences between raters).

APPENDIX 2: DOC TASK FORCE RATING SCHEME FOR IC COEFFICIENTS

IC Rating	Cronbach α (4–6 Item Scale)	Cronbach α (7+ Item Scale)
Unacceptable	<.60	<.70
Acceptable	.60–.69	.70–.79
Good	.70–.79	.80–.89
Excellent	.80+	.90+

APPENDIX 3: DOC TASK FORCE RATING SCHEME FOR IRR AND TRR COEFFICIENTS

IRR and TRR Ratings*	Coefficient†
Unacceptable	<.70
Acceptable	.70–.79
Good	.80–.89
Excellent	.90+

*If only a correlation coefficient is reported, then the qualifier "(systematic error not evaluated)" is noted after rating.
†Includes Cohen κ, ICC, and Pearson/Spearman correlation coefficients.

APPENDIX 4: AAN EVIDENCE CLASSIFICATION SCHEME FOR RATING DOC DIAGNOSTIC AND PROGNOSTIC METHODOLOGIES

Class	Diagnostic Study Methodologic Features	Prognostic Study Methodologic Features
I	Low risk of bias. A prospective study of a broad sample suspected of having DOC using a reference standard (eg, consensus-based diagnosis or use of a standardized scale) for case definition of true diagnosis or scale score. The reference standard has less than 70% item content overlap with the experimental scale. The reference standard and experimental scale are administered using a blind evaluation of the entire sample.	Low risk of bias. A prospective study of a broad DOC sample that uses a DOC scale as a baseline predictor. A reference standard measures future risk (eg, PVS, very severe disability) or positive outcome (eg, emergence from MCS, functional capacity). The outcome measure has less than 70% content overlap with the DOC scale and is administered blind to the DOC scale (predictor) results. At least 80% of the initial sample completes the outcome measure.
II	Moderate risk of bias. A prospective study of a narrow sample suspected of having DOC OR a retrospective study of a broad DOC sample using a reference standard (eg, consensus based diagnosis or use of a standardized assessment scale) for case definition of true diagnosis or scale score. The content of the reference standard has less than 70% item content overlap with the experimental scale. The experimental scale and reference standard are administered using a blind evaluation of the entire sample.	Moderate risk of bias. A prospective study of a narrow DOC sample OR a retrospective study of a broad DOC sample that uses a DOC scale as a baseline predictor. A reference standard measures future risk (eg, PVS, very severe disability) or positive outcome (eg, emergence from MCS, functional capacity). The outcome measure has less than 70% content overlap with the DOC scale and is administered blind to the DOC scale (predictor) results. At least 80% of the initial sample completes the outcome measure.
III	Moderate to high risk of bias. A retrospective study of a narrow sample OR the content of the reference standard has ≥70% item content overlap with the experimental scale OR independent but unblind raters administer the experimental scale and reference standard OR subsets of persons are systematically removed from study (eg, only high and low scorers are analyzed).	Moderate to high risk of bias. A retrospective study of a narrow DOC sample OR the content of the outcome measure has ≥70% content overlap with the DOC scale OR independent but unblind raters administer the outcome scale OR subsets of persons are systematically removed from study (eg, only high and low scorers are analyzed) OR ≤80% of the initial sample completes the outcome measure.

APPENDIX 4: AAN EVIDENCE CLASSIFICATION SCHEME FOR RATING DOC DIAGNOSTIC AND PROGNOSTIC METHODOLOGIES (Cont'd)

Class	Diagnostic Study Methodologic Features	Prognostic Study Methodologic Features
IV	Very high risk of bias. Any design in which the entire sample has the same diagnosis (eg, all VS) OR no reference standard is administered OR the same rater administers both the experimental scale and the reference standard.	Very high risk of bias. Any design in which the entire sample has the same diagnosis (eg, all VS) or outcome (eg, all emerged from MCS) OR no reference standard (outcome measure) is administered OR the same rater administers both the predictor scale and the outcome measure.

References

1. Laureys S, Perrin F, Schnakers C, Boly M, Majerus S. Residual cognitive function in comatose, vegetative and minimally conscious states. Curr Opin Neurol 2005;18:726-33.
2. Giacino J, Smart C. Recent advances in behavioral assessment of individuals with disorders of consciousness. Curr Opin Neurol 2007;20:614-9.
3. Childs NL, Mercer WN, Child HW. Accuracy of diagnosis of persistent vegetative state. Neurology 1993;43:1465-7.
4. Andrews K, Murphy L, Munday R, Littlewood C. Misdiagnosis of the vegetative state: retrospective study in a rehabilitation unit. BMJ 1996;313:13-6.
5. Schnakers C, Vanhaudenhuyse A, Giacino J, et al. Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment. BMC Neurol 2009;9:35.
6. Kotchoubey B, Lang S, Mezger G, et al. Information processing in severe disorders of consciousness: vegetative state and minimally conscious state. Clin Neurophysiol 2005;116:2441-53.
7. Laureys S, Giacino J, Schiff N, Schabus M, Owen A. How should functional imaging of patients with disorders of consciousness contribute to their clinical rehabilitation needs? Curr Opin Neurol 2006;19:520-7.
8. Perrin F, Schnakers C, Degueldre C, et al. Brain response to one's own name in vegetative state, minimally conscious state and locked-in syndrome. Arch Neurol 2006;63:562-9.
9. Giacino J, Hirsch J, Schiff N, Laureys S. Functional neuroimaging applications for assessment and rehabilitation planning in patients with disorders of consciousness. Arch Phys Med Rehabil 2006;87(Suppl 2):S67-76.
10. Owen A, Coleman M. Functional neuroimaging of the vegetative state. Nat Rev Neurosci 2008;9:235-43.
11. Schnakers C, Perrin F, Schabus M, et al. Voluntary brain processing in disorders of consciousness. Neurology 2008;71: 1614-20.
12. Plum F, Posner J. The pathologic physiology of signs and symptoms of coma: the diagnosis of stupor and coma. 3rd ed. Philadelphia: FA Davis; 1982.
13. American Academy of Neurology Quality Standards Subcommittee. Practice parameter: assessment and management of persons in the persistent vegetative state. Neurology 1995;45: 1015-8.
14. Giacino JT, Ashwal S, Childs N, et al. The minimally conscious state: definition and diagnostic criteria. Neurology 2002;58:349-53.
15. Plum F. Coma and related global disturbances of the human conscious state. In: Jones E, Peters P, editors. Cerebral cortex. Vol 9. New York: Plenum Pr; 1991.
16. Multi-Society Task Force Report on PVS. Medical aspects of the persistent vegetative state. N Engl J Med 1994;330:1499-508, 1572-9.
17. Jennett B, Plum F. Persistent vegetative state after brain damage: a syndrome in search of a name. Lancet 1972;1:734-7.
18. Giacino JT, Kalmar K. The vegetative and minimally conscious states: a comparison of clinical features and functional outcome. J Head Trauma Rehabil 1997;12:36-51.
19. Giacino JT, Zasler ND. Outcome after severe traumatic brain injury: coma, the vegetative state, and the minimally responsive state. J Head Trauma Rehabil 1995;10:40-56.
20. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. Lancet 1974;2:81-4.
21. Ansell BJ, Keenan JE. The Western Neuro Sensory Stimulation Profile: a tool for assessing slow-to-recover head-injured patients. Arch Phys Med Rehabil 1989;70:104-8.
22. Giacino JT, Kezmarsky MA, DeLuca J, Cicerone KD. Monitoring rate of recovery to predict outcome in minimally responsive patients. Arch Phys Med Rehabil 1991;72:897-901.
23. Rappaport M, Dougherty AM, Kelting DL. Evaluation of coma and vegetative states. Arch Phys Med Rehabil 1992;73: 628-34.
24. Edlund W, Gronseth G, So Y, Franklin G. Clinical practice guideline process manual. St Paul: American Academy of Neurology; 2004.
25. Mayer SA, Dennis LJ, Peery S, et al. Quantification of lethargy in the neuro-ICU: the 60-second test. Neurology 2003;61:543-5.
26. Freeman EA. The Coma Exit Chart: assessing the patient in prolonged coma and the vegetative state. Brain Inj 1996;10: 615-24.
27. Freeman EA. The clinical assessment of coma. Neuropsychol Rehabil 1993;3:139-147.
28. Freeman EA. Protocols for the vegetative state. Brain Inj 1997; 11:837-49.
29. Gruner ML, Terhaag D. Multimodal early onset stimulation (MEOS) in rehabilitation after brain injury. Brain Inj 2000;14: 585-94.
30. Lippert-Gruner M, Wedekind C, Ernestus RI, Klug N. Early rehabilitative concepts in therapy of the comatose brain injured patients. Acta Neurochir Suppl 2002;79:21-3.
31. Giacino JT, Kalmar K. Diagnostic and prognostic guidelines for the vegetative and minimally conscious states. Neuropsychol Rehabil 2005;15:166-74.
32. Kalmar K, Giacino JT. The JFK Coma Recovery Scale—Revised. Neuropsychol Rehabil 2005;15:454-60.
33. Pilon M, Sullivan SJ. Motor profile of patients in minimally responsive and persistent vegetative states. Brain Inj 1996;10: 421-37.
34. Rappaport M. The Disability Rating and Coma/Near-Coma scales in evaluating severe head injury. Neuropsychol Rehabil 2005;15:442-53.
35. Pape TL, Senno RG, Guernon A, Kelly JP. A measure of neurobehavioral functioning after coma. Part II: clinical and scientific implementation. J Rehabil Res Dev 2005;42:19-27.
36. Ohta T. Phenomenological aspects of consciousness—its disturbance in acute and chronic stages. Acta Neurochir 2005;93: 191-3.
37. Wijdicks EF. Clinical scales for comatose patients: the Glasgow Coma Scale in historical context and the new FOUR score. Rev Neurol Dis 2006;3:109-17.

38. Benzer A, Mitterschiffthaler G, Marosi M, et al. Prediction of non-survival after trauma: Innsbruck Coma Scale. Lancet 1991; 338:977-8.
39. Benzer A, Traweger C, Ofner D, Marosi M, Luef G, Schmutzhard E. Statistical modelling in analysis of outcome after trauma Glasgow-Coma-Scale and Innsbruck-Coma-Scale. Anasthesiol Intensivmed Notfallmed Schmerzther 1995;30: 231-5.
40. Neumann N, Kotchoubey B. Assessment of cognitive functions in severely paralysed and severely brain-damaged patients: neuropsychological and electrophysiological methods. Brain Res Brain Res Protoc 2004;14:25-36.
41. Cossa FM, Fabiani M, Farinato A, Laiacona M, Capitani E. The "preliminary neuropsychological battery": an instrument to grade the cognitive level of minimally responsive patients. Brain Inj 1999;13:583-92.
42. Noda R, Maeda Y, Yoshino A. Effects of musicokinetic therapy and spinal cord stimulation on patients in a persistent vegetative state. Acta Neurochir Suppl 2003;87:23-6.
43. Noda R, Maeda Y, Yoshino A. Therapeutic time window for musicokinetic therapy in a persistent vegetative state after severe brain damage. Brain Inj 2004;18:509-15.
44. Stalhammar D, Starmark JE. Assessment of responsiveness in head injury patients: the Glasgow Coma Scale and some comments on alternative methods. Acta Neurochir 1986;36:91-4.
45. Starmark JE, Lindgren S. Possible mechanisms of "vegetative state." Acta Neurochir Suppl 1986;36:121-2.
46. Starmark JE, Lindgren S. Is it possible to define a general "conscious level"? Acta Neurochir 1986;36:103-5.
47. Starmark JE, Stilhammer D, et al. Assessment of responsiveness: a comparison between the Glasgow Coma Scale (GCS) and the Reaction Level Scale (RLS-85). Acta Neurochir Wien 1987;84: 150-51.
48. Stein SC, Spettell C, Young G, Ross SE. Limitations of neurological assessment in mild head injury. Brain Inj 1993;7:425-30.
49. Walther SM, Jonasson U, Gill H. Comparison of the Glasgow Coma Scale and the Reaction Level Scale for assessment of cerebral responsiveness in the critically ill. Intensive Care Med 2003;29:933-8.
50. Starmark JE, Stålhammar D, Holmgren E. The Reaction Level Scale (RLS85): manual and guidelines. Acta Neurochir 1988;91: 12-20.
51. Wilson SL, Brock D, Powell GE, Thwaites H, Elliott K. Constructing arousal profiles for vegetative state patients—a preliminary report. Brain Inj 1996;10:105-13.
52. Wilson SL. Magnetic-resonance imaging and prediction of recovery from post-traumatic vegetative state. Lancet 1998;352: 485.
53. Gill-Thwaites H, Munday R. The Sensory Modality Assessment and Rehabilitation Technique (SMART): a comprehensive integrated assessment and treatment protocol for the vegetative state and minimally responsive patient. Neuropsychol Rehabil 1999; 9:305-20.
54. Majerus S, Gill-Thwaites H, Andrews K, Laureys S. Behavioral evaluation of consciousness in severe brain damage. Prog Brain Res 2005;150:397-413.
55. Wilson FC, Graham LE, Watson T. Vegetative and minimally conscious states: serial assessment approaches in diagnosis and management. Neuropsychol Rehabil 2005;15:431-41.
56. Hall ME, MacDonald S, Young GC. The effectiveness of directed multisensory stimulation versus non-directed stimulation in comatose CHI patients: pilot study of a single subject design. Brain Inj 1992;6:435-45.
57. Davis AE, Gimenez A. Cognitive-behavioral recovery in comatose patients following auditory sensory stimulation. J Neurosci Nurs 2003;35:202-9, 14.
58. Berger E, Wörgötter G, Oppolzer A, Kessler J, Vavrik K, Fiala S. Neurological rehabilitation in children and adolescents. Pediatr Rehabil 1997;1:229-33.
59. Berger E, Vavrik K, Hochgatterer P. Vigilance scoring in children with acquired brain injury: Vienna Vigilance Score in comparison with usual coma scales. J Child Neurol 2001; 16:236-40.
60. Horn S, Shiel A, McLellan L, et al. A review of behavioural assessment scales for monitoring recovery in and after coma with pilot data on a new scale of visual awareness. Neuropsychol Rehabil 1993;3:121-37.
61. Wilson FC, Harpur J, Watson T, Morrow JI. Vegetative state and minimally responsive patients—regional survey, long-term case outcomes and service recommendations. NeuroRehabilitation 2002;17:231-6.
62. Shiel A, Wilson BA. Can behaviours observed in the early stages of recovery after traumatic brain injury predict poor outcome? Neuropsychol Rehabil 2005;15:494-502.
63. Wilson FC, Elder V, McCrudden E, Caldwell S. Analysis of Wessex Head Injury Matrix (WHIM) scores in consecutive vegetative and minimally conscious state patients. Neuropsychol Rehabil 2009;19:754-60.
64. Ansell BJ. Slow-to-recover brain-injured patients: rationale for treatment. J Speech Hear Res 1991;34:1017-22.
65. Keenan JE. Assessment tools for severely head-injured adults. Cogn Rehabil 1989;7:24-6.
66. Lammi MH, Smith VH, Tate RL, Taylor CM. The minimally conscious state and recovery potential: a follow-up study 2 to 5 years after traumatic brain injury. Arch Phys Med Rehabil 2005; 86:746-54.
67. Patrick PD, Buck ML, Conaway MR, Blackman JA. The use of dopamine enhancing medications with children in low response states following brain injury. Brain Inj 2003;17:497-506.
68. Taylor CM, Aird VH, Tate RL, Lammi MH. Sequence of recovery during the course of emergence from the minimally conscious state. Arch Phys Med Rehabil 2007;88:521-5.
69. O'Dell MW, Jasin P, Lyons N, Stivers M, Meszaro F. Standardized assessment instruments for minimally-responsive, brain-injured patients. NeuroRehabilitation 1996;6:45-55.
70. O'Dell MW, Jasin P, Stivers M, Lyons N, Schmidt S, Moore DE. Interrater reliability of the Coma Recovery Scale. J Head Trauma Rehabil 1996;11:61-6.
71. Schnakers C, Majerus S, Laureys S. Bispectral analysis of electroencephalogram signals during recovery from coma: preliminary findings. Neuropsychol Rehabil 2005;15:381-8.
72. Giacino JT, Kalmar K, Whyte J. The JFK Coma Recovery Scale-Revised: measurement characteristics and diagnostic utility. Arch Phys Med Rehabil 2004;85:2020-9.
73. Schnakers C, Majerus S, Giacino J, et al. A French validation study of the Coma Recovery Scale-Revised. Brain Inj 2008;22: 786-92.
74. Talbot LR, Whitaker HA. Brain-injured persons in an altered state of consciousness: measures and intervention strategies. Brain Inj 1994;8:689-99.
75. Stanczak DE, White JG 3rd, Gouview WD, et al. Assessment of level of consciousness following severe neurological insult: a comparison of the psychometric qualities of the Glasgow Coma Scale and the Comprehensive Level of Consciousness Scale. J Neurosurg 1984;60:955-60.
76. Johnston MD, Thomas L, Stanczak DE. Construct validity of the Comprehensive Level of Consciousness Scale: a comparison of behavioral and neurodiagnostic measures. Arch Clin Neuropsychol 1996;11:703-11.
77. Pape TLB, Heinemann AW, Kelly JP, Hurder AG, Lundgren S. A measure of neurobehavioral functioning after coma. Part I:

theory, reliability, and validity of Disorders of Consciousness Scale. J Rehabil Res Dev 2005;42:1-17.
78. Pape TL, Lundgren S, Heinemann AW, et al. Establishing a prognosis for functional outcome during coma recovery. Brain Inj 2006;20:743-58.
79. Pape TL, Tang C, Guernon A, et al. Predictive value of the Disorders of Consciousness Scale (DOCS). PM R 2009;1: 152-61.
80. Wijdicks EF, Bamlet WR, Maramattom BV, Manno EM, McClelland RL. Validation of a new coma scale: the FOUR score. Ann Neurol 2005;58:585-93.
81. Wolf CA, Wijdicks EFM, Bamlet WR, McClelland RL. Further validation of the FOUR score coma scale by intensive care nurses. Mayo Clin Proc 2007;82:435-8.
82. Stead LG, Wijdicks EFM, Bhagra A, et al. Validation of a new coma scale, the FOUR score, in the emergency department. Neurocrit Care 2009;10:50-4.
83. Eken C, Kartal M, Bucanli A, Eray O. Comparison of the Full Outline of Unresponsiveness Score Coma Scale and the Glasgow Coma Scale in an emergency setting population. Eur J Emerg Med 2009;16:29-36.
84. Mukherjee KK, Sharma BS, Ramanathan SM, Khandelwal N, Kak VK. A mathematical outcome prediction model in severe head injury: a pilot study. Neurol India 2000;48:43-8.
85. Born JD. The Glasgow-Liege Scale: prognostic value and evolution of motor response and brain stem reflexes after severe head injury. Acta Neurochir 1988;91:1-11.
86. Born JD, Hans P, Albert A, Bonnal J. Interobserver agreement in assessment of motor response and brain stem reflexes. Neurosurgery 1987;20:513-7.
87. Born JD, Albert A, Hans P, Bonnal J. Relative prognostic value of best motor response and brain stem reflexes in patients with severe head injury. Neurosurgery 1985;16:595-601.
88. Diringer MN, Edwards DF. Does modification of the Innsbruck and the Glasgow Coma Scales improve their ability to predict functional outcome? Arch Neurol 1997;54:606-11.
89. Berek K, Schinnerl A, Traweger C, Lechleitner P, Baubin M, Aichner F. The prognostic significance of coma-rating, duration of anoxia and cardiopulmonary resuscitation in out-of hospital cardiac arrest. J Neurol 1997;244:556-61.
90. Borer-Alafi N, Gil M, Sazbon L, Korn C. Loewenstein communication scale for the minimally responsive patient. Brain Inj 2002;16:593-609.
91. Johnstone AJ, Lohlun JC, Miller JD, et al. A comparison of the Glasgow Coma Scale and the Swedish Reaction Level Scale. Brain Inj 1993;7:501-6.
92. Matousek M, Takeuchi E, Starmark JE, Stalhammar D. Quantitative EEG analysis as a supplement to the clinical coma scale RLS85. Acta Anaesthesiol Scand 1996;40:824-31.
93. Stålhammar D, Starmark JE, Holmgren E, et al. Assessment of responsiveness in acute cerebral disorders: a multicentre study on the reaction level scale (RLS 85). Acta Neurochir 1988;90:73-80.
94. Starmark JE, Stålhammar D, Holmgren E, Rosander B. A comparison of the Glasgow Coma Scale and the Reaction Level Scale (RLS85). J Neurosurg 1988;69:699-706.
95. Tesseris J, Pantazidis N, Routsi C, Fragoulakis D. A comparative study of the Reaction Level Scale (RLS85) with Glasgow Coma Scale (GCS) and Edinburgh-2 Coma Scale (modified) (E2CS(M)). Acta Neurochir 1991;110:65-76.
96. Wilson SL, Powell GE, Brock D, Thwaites H. Behavioural differences between patients who emerged from vegetative state and those who did not. Brain Inj 1996;10:509-16.
97. Wilson SL, Powell GE, Brock D, Thwaites H. Vegetative state and responses to sensory stimulation: an analysis of 24 cases. Brain Inj 1996;10:807-18.
98. Gill-Thwaites H. The Sensory Modality Assessment Rehabilitation Technique—a tool for assessment and treatment of patients with severe brain injury in a vegetative state. Brain Inj 1997;11: 723-34.
99. Wilson SL, Gill-Thwaites H. Early indication of emergence from vegetative state derived from assessments with the SMART—a preliminary report. Brain Inj 2000;14:319-31.
100. Gill-Thwaites H, Munday R. The Sensory Modality Assessment and Rehabilitation Technique (SMART): a valid and reliable assessment for vegetative state and minimally conscious state patients. Brain Inj 2004;18:1255-69.
101. Rader MA, Alston JB, Ellis DW. Sensory stimulation of severely brain-injured patients. Brain Inj 1989;3:141-7.
102. Rader MA, Ellis DW. The Sensory Stimulation Assessment Measure (SSAM): a tool for early evaluation of severely brain-injured patients. Brain Inj 1994;8:309-21.
103. Shiel A, Horn SA, Wilson BA, Watson MJ, Campbell MJ, McLellan DL. The Wessex Head Injury Matrix (WHIM) main scale: a preliminary report on a scale to assess and monitor patient recovery after severe head injury. Clin Rehabil 2000;14:408-16.
104. Majerus S, Van der Linden M, Shiel A. Wessex Head Injury Matrix and Glasgow/Glasgow–Liége Coma Scale: a validation and comparison study. Neuropsychol Rehabil 2000;10:167-84.
105. Ansell BJ. Slow-to-recover patients: improvement to rehabilitation readiness. J Head Trauma Rehabil 1993;8:88-98.
106. Ansell BJ. Visual tracking behavior in low functioning head-injured adults. Arch Phys Med Rehabil 1995;76:726-31.
107. Anastasi A. Psychological testing. 6th ed. New York:MacMillan Publishing; 1988.
108. Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil 2000;81(Suppl 2):S15-20.
109. Bédard M, Martin NJ, Krueger P, Brazil K. Assessing reproducibility of data obtained with instruments based on continuous measurements. Exp Aging Res 2000;26:353-65.
110. Biddle AK, Watson LR, Hooper CR, Lohr KN, Sutton SF. Criteria for determining disability in speech-language disorders. AHRQ Evid Rep Technol Assess 2002;52:1-60.
111. Bland JM, Altman DG. Cronbach's alpha. BMJ (Clinical Res Ed) 1997;314:572.
112. Bonnet DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med 2002;21:1331-5.
113. Carmines EG, Zeller, RA. Reliability and validity. London: Sage Publications; 1980.
114. Cohen J, Cohen P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale: Lawrence Erlbaum Associates; 1989.
115. Crocker L, Algina J. Introduction to classical and modern test theory. Orlando: Harcourt Brace; 1986.
116. DeVellis RF. Scale development: theory and applications. 2nd ed. Applied Social Research Methods Series. Vol 26. Thousand Oaks: Sage Publications; 2003.
117. Fleiss J, Cohen JA. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613-9.
118. Fleiss JL. Statistical methods for rates and proportions. New York: Wiley; 1981.
119. Fleiss JL. Design and analysis of clinical experiments. New York: Wiley; 1986.
120. Frank-Stromborg M, Olsen S. Instruments for clinical health-care research. 3rd ed. Sudbury: Jones and Bartlett Publishers; 2004.
121. Lexell JE, Downham DY. How to assess the reliability of measurements in rehabilitation. Am J Phys Med Rehabil 2005;84:719-23.
122. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.

123. Raykov T. Equivalent structural equation models and group equality constraints. Multivariate behavioral research. 1997;32:95-105.
124. Shoukri MM. Sample size requirements for the design of reliability study: review and new results. Stat Methods Med Res 2004;13:251-71.
125. Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res 1998;7:301-17.
126. Cameron J, Gignac M. "Timing is Right": a conceptual framework for addressing the support needs of family caregivers to stroke survivors from hospital to the home. Patient Educ Couns 2008;70:305-14.
127. Hillier S, Metzer J. Awareness and perceptions of outcomes after traumatic brain injury. Brain Inj 1997;11:525-36.
128. Glasgow R, Magid D, Beck A, Ritzqoller D, Estabrooks P. Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care 2005; 43:551-7.
129. Tunis S, Stryer D, Clancey C. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003;290:1624-32.
130. Whyte J. Clinical trials in rehabilitation: what are the obstacles? Am J Phys Med Rehabil 2003;82(Suppl 10):S16-21.
131. Hahn E, Cella D. Health outcomes assessment in vulnerable populations: measurement challenges and recommendations. Arch Phys Med Rehabil 2003;84(Suppl 2):S35-42.