
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 pair 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 SYNTHESI S: 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...
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. 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. 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.

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, the VS, and the MCS. Coma represents a state of un arousable unresponsiveness in which there is no evidence of self-awareness or environmental awareness. The eyes remain continuously closed, purposeful awareness. Conscious behaviors are often sub- stantially 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.

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. 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.

© 2010 by the American Congress of Rehabilitation Medicine

List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAN</td>
<td>American Academy of Neurology</td>
</tr>
<tr>
<td>ACRM</td>
<td>American Congress of Rehabilitation Medicine</td>
</tr>
<tr>
<td>BI-ISIG</td>
<td>Brain Injury-Interdisciplinary Special Interest Group</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CLOCS</td>
<td>Comprehensive Levels of Consciousness Scale</td>
</tr>
<tr>
<td>CNC</td>
<td>Coma/Near-Coma Scale</td>
</tr>
<tr>
<td>CPC</td>
<td>Clinical Practice Committee</td>
</tr>
<tr>
<td>CRS-R</td>
<td>Coma Recovery Scale-Revised</td>
</tr>
<tr>
<td>DOC</td>
<td>disorders of consciousness</td>
</tr>
<tr>
<td>DOCS</td>
<td>Disorders of Consciousness Scale</td>
</tr>
<tr>
<td>DRS</td>
<td>Disability Rating Scale</td>
</tr>
<tr>
<td>FOUR</td>
<td>Full Outline of UnResponsiveness Score</td>
</tr>
<tr>
<td>GCS</td>
<td>Glasgow Coma Scale</td>
</tr>
<tr>
<td>GLS</td>
<td>Glasgow-Liege Coma Scale</td>
</tr>
<tr>
<td>IC</td>
<td>internal consistency</td>
</tr>
<tr>
<td>ICC</td>
<td>intraclass correlation coefficient</td>
</tr>
<tr>
<td>INNS</td>
<td>Innsbruck Coma Scale</td>
</tr>
<tr>
<td>IRR</td>
<td>interrater reliability</td>
</tr>
<tr>
<td>LOEW</td>
<td>Loewenstein Communication Scale</td>
</tr>
<tr>
<td>MCS</td>
<td>minimally conscious state</td>
</tr>
<tr>
<td>MSKTC</td>
<td>Model Systems Knowledge Translation Center</td>
</tr>
<tr>
<td>NIDRR</td>
<td>National Institute on Disability and Rehabilitation Research</td>
</tr>
<tr>
<td>PVS</td>
<td>persistent vegetative state</td>
</tr>
<tr>
<td>RLSBS</td>
<td>Swedish Reaction Level Scale-1985</td>
</tr>
<tr>
<td>SMART</td>
<td>Sensory Modality Assessment Technique</td>
</tr>
<tr>
<td>SSAM</td>
<td>Sensory Stimulation Assessment Measure</td>
</tr>
<tr>
<td>TRR</td>
<td>test–retest reliability</td>
</tr>
<tr>
<td>VS</td>
<td>vegetative state</td>
</tr>
<tr>
<td>WHIM</td>
<td>Wessex Head Injury Matrix</td>
</tr>
<tr>
<td>WNSSP</td>
<td>Western Neuro Sensory Stimulation Profile</td>
</tr>
</tbody>
</table>
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. 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, 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. 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 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, 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 non-traumatic 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 and an additional 4 were excluded during
### Table 1: Description of Disorders of Consciousness Scales

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

*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 proceeded as follows: 22,69–71 Thus, guideline development was based on 37 articles5,21,23,72–106 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
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). 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 articles, discussed disagreements, and reached consensus. Only the CRS-R, SMART, CNC, and RLS85 provided applied. Only the CRS-R, SMART, CNC, and RLS85 provided diagnostic guidelines 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.

**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.60,61,116,122 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.60,61 One class I study (N=57) of the 32-item WNSSP showed a Cronbach alpha of .95.21 One class I study (N=80) of the 7-item SMART found a Cronbach alpha of .88, whereas 1 class I study (N=80) of the 23-item CRSc-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 I/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 I/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.73 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**

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

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 coefficient: unacceptable <.70, acceptable .70–.79, good = .80–.89, excellent ≥ .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 ≥.80 are considered excellent.
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 .85,130 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 \(\kappa=.89\)).82 Two small sample class II/III studies of the CRS-R by using different methodologies and with different raters reported a \(\kappa\) of .80.83–85 One small sample class II/III study of the LOEW by using a single administration of scale items with 2 independent raters reported a \(\kappa\) of .98.96 Two class II/III studies of the RLS85 by using the same methodology reported IRRs just above and below the cutoff score for acceptability (\(\kappa=.69\) 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 (\(\kappa=.69\)) and motor ratings (\(\kappa=.85\)).96–98 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 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 administrations
ratings 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.\(^\text{72}\)

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.\(^\text{100}\)

Single class IV studies of the CLOCS and SSAM reported good to excellent correlation values but ambiguities with regard to independence of ratings and how much time elapsed between scale administrations introduced a high risk of bias.\(^\text{103}\)

No TTR 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 TTR (class II/III). TTR 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 produced a Pearson correlation of .75 (95% CI, .65–.82).\(^\text{75}\)

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).\(^\text{84}\)

Nine of the 11 class IV studies reported very high correlations (absolute coefficient value $\geq .69$) 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,\(^\text{72,73}\) CNC,\(^\text{23,74}\) WNSSP,\(^\text{21}\) RLS85,\(^\text{84}\) and SMART,\(^\text{104}\) the same rater completed the investigational scale and a reference standard. In 3 class IV studies involving the CLOCS,\(^\text{76}\) SSAM,\(^\text{101,102}\) and WHIM,\(^\text{104}\) masking
status was not stated. Other research design issues in class IV studies included small samples, samples with narrow DOC representation, and use of reference standards with item content that highly overlapped (>70%) the investigational scales. 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 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 hierarchical structure of the WHIM.

Conclusions. The DOCS possibly has unidimensional hierarchical interval characteristics when used with persons with closed head injury (class III). The unidimensional hierarchical structure of the DOCS in other acquired brain injury subgroups has not been adequately established (class IV). The unidimensional hierarchical 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 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 and RLS85 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 survivors’ 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, to acute care period in 4 studies, to subacute and chronic periods up to 5 years postinjury in 4 studies. With regard to outcome measures, 6 studies used the Glasgow Outcome Scale, 4 studies dichotomized Rankin Scale scores, and other studies used dichotomized recovery of consciousness and rehabilitation readiness, or discharge settings. The timing of outcome measurement greatly varied.
Table 6: Methodologic Issues and Statistical Outcomes in Studies With Prognostic Validity Analyses for DOC Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Author (y)</th>
<th>Evidence Class</th>
<th>Method Issues</th>
<th>Predictor Score</th>
<th>Timing of Pred Mess</th>
<th>Ref Std Outcome</th>
<th>Timing of FUP Mess</th>
<th>Stats Calc</th>
<th>Coefficient (95% CI)</th>
<th>OddsR (95% CI)</th>
<th>% Correct</th>
<th>Sens (95% CI)</th>
<th>Spec (95% CI)</th>
<th>PPV (95% CI)</th>
<th>NPV (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOUR</td>
<td>Wolf (2007)</td>
<td>I</td>
<td>Total</td>
<td>≥24h</td>
<td>Ranlin Good 30d</td>
<td>LR</td>
<td>.58 (0.41–0.82)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>FOUR</td>
<td>Stead (2009)</td>
<td>III</td>
<td>Total ≥9</td>
<td>≥24h</td>
<td>Ranlin Good Acute DC</td>
<td>LR</td>
<td>.43 (0.26–0.71)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>FOUR</td>
<td>Wijdicks (2005)</td>
<td>IV</td>
<td>Total ICU stay</td>
<td>≥9</td>
<td>Ranlin Good 3mo</td>
<td>LR</td>
<td>.84 (0.77–0.92)</td>
<td>.75 (NS)</td>
<td>.76 (NS)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>FOUR</td>
<td>Eken (2009)</td>
<td>IV</td>
<td>Total ICU stay</td>
<td>≥9</td>
<td>Ranlin Good 3mo</td>
<td>AUC</td>
<td>.75 (0.68–0.81)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>INNS</td>
<td>Diringer (1997)</td>
<td>I</td>
<td>Total (minus Oral Autonim Score)</td>
<td>≥24h</td>
<td>Death 3mo (post-DC)</td>
<td>WL</td>
<td>.72†</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>GLS</td>
<td>Born (1985)</td>
<td>III</td>
<td>Total = age</td>
<td>≥24h</td>
<td>Independent GOS Good Mod</td>
<td>6mo</td>
<td>LR</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>GLS</td>
<td>Born (1988)</td>
<td>III</td>
<td>Motor ref lex</td>
<td>≤24h</td>
<td>Independent GOS Good Mod</td>
<td>6mo</td>
<td>LR</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>GLS</td>
<td>Mukherjee (2000)</td>
<td>IV</td>
<td>M Total</td>
<td>≥24h</td>
<td>Not conscious Acute DC</td>
<td>MR</td>
<td>NS</td>
<td>84% (all groups)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>DOCS</td>
<td>Pape (2006)</td>
<td>IV</td>
<td>M, F Total ≤48</td>
<td>≤48d</td>
<td>Carroll Chart 1y</td>
<td>CHART scales 21–365d</td>
<td>1y</td>
<td>LR</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>DOCS</td>
<td>Pape (2009)</td>
<td>IV</td>
<td>M, F DOCS change ≤4</td>
<td>≤48d</td>
<td>Carroll Chart 1y</td>
<td>CHART scales 21–365d</td>
<td>1y</td>
<td>LR</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>LOEW</td>
<td>Borer-Alafi (2002)</td>
<td>IV</td>
<td>M, F Total ≤50</td>
<td>≤50</td>
<td>Not rehab ready 26–632d</td>
<td>LR</td>
<td>1.12 (1.04–1.20)</td>
<td>.91 (NS)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>RLSRS</td>
<td>Johnstone (1993)</td>
<td>IV</td>
<td>M Level 4–8</td>
<td>≤4–8</td>
<td>Acute admission</td>
<td>GOS Good</td>
<td>Acute DC</td>
<td>.96 (0.91–0.98)</td>
<td>.81 (1.17–.82)</td>
<td>.81 (0.75–0.86)</td>
<td>.88 (0.45–.85)</td>
<td>.88 (0.45–.85)</td>
<td>.88 (0.45–.85)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLOCS</td>
<td>Stanczak (1994)</td>
<td>IV</td>
<td>M, N, F Level &lt;3d</td>
<td>&lt;3d</td>
<td>Modif. GOS</td>
<td>Acute DC</td>
<td>MR</td>
<td>.32 (NS)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>CLOCS</td>
<td>Johnston (1996)</td>
<td>IV</td>
<td>M, N, F Level &gt;=5 or transformed</td>
<td>&lt;3d</td>
<td>Modif. GOS</td>
<td>Acute DC</td>
<td>MR</td>
<td>.07† (NS)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

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 scales, 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%; 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, severely moderate disability, severe disability, or death; MR=multiple regression, WL=Wilkes; Timing of FUP Mess, timing of the outcome measurement relative to the injury date unless noted; Timing of Pred Mess, 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 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 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 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 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. 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.

Commonly encountered research design issues in the 13 class IV studies included lack of masking between predictor and outcome ratings in 12 studies; greater than 20% loss to follow-up in 6 studies; and a narrow sample composition in 3 studies. 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. 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 WNSSP 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

<table>
<thead>
<tr>
<th>Scale</th>
<th>Q1 Standardized Admin/Scoring</th>
<th>Q2 Content Validity</th>
<th>Q3a IC</th>
<th>Q3b IRR</th>
<th>Q3c TRR</th>
<th>Q4 &amp; Q5 Criterion Validity</th>
<th>Q6 Diagnostic Validity</th>
<th>Q7 Prognostic Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRS-R</td>
<td>Acceptable</td>
<td>Excellent</td>
<td>Good</td>
<td>Good</td>
<td>Excellent (class IV)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
</tr>
<tr>
<td>SMART</td>
<td>Acceptable</td>
<td>Good</td>
<td>NA</td>
<td>Excellent (class IV)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>WNSSP</td>
<td>Acceptable</td>
<td>Good</td>
<td>Excellent (class I)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>SSAM</td>
<td>Acceptable</td>
<td>Good</td>
<td>Unproven (not studied)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>WHIM</td>
<td>Acceptable</td>
<td>Good</td>
<td>Unproven (not studied)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>DOCS</td>
<td>Acceptable</td>
<td>Acceptable</td>
<td>Good</td>
<td>Unproven (class IV)</td>
<td>Unproven</td>
<td>Construct valid*</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>CNC</td>
<td>Acceptable</td>
<td>Acceptable</td>
<td>Unacceptable (class IV)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>CLOCS</td>
<td>Unacceptable</td>
<td>Acceptable</td>
<td>Good</td>
<td>Unproven (class IV)</td>
<td>Unproven</td>
<td>Strong</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>LOEW</td>
<td>Unacceptable</td>
<td>Acceptable</td>
<td>Unproven (not studied)</td>
<td>Excellent (class II/III)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>RLS85</td>
<td>Unacceptable</td>
<td>Acceptable</td>
<td>NA</td>
<td>Unproven (class IV)</td>
<td>Unproven</td>
<td>Strong</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>FOUR</td>
<td>Unacceptable</td>
<td>Unacceptable</td>
<td>Excellent (multiple class I)</td>
<td>Good</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>INNS</td>
<td>Unacceptable</td>
<td>Unacceptable</td>
<td>Acceptable (class I)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
<tr>
<td>GLS</td>
<td>Unacceptable</td>
<td>Unacceptable</td>
<td>Unproven (not studied)</td>
<td>Unacceptable (class II/III)</td>
<td>Unproven</td>
<td>Unproven</td>
<td>Unproven</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: multiple class I, established as reliable, diagnostic valid, or prognostic valid; class I study or multiple class II or III studies, probably reliable, diagnostic valid, or prognostic valid; class 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 study72 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.26-131

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 assessments 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 3-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.

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

<table>
<thead>
<tr>
<th>Class</th>
<th>IRR and TRR Methodologic Features Associated With Each Evidence Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>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 $\alpha$) is calculated.</td>
</tr>
<tr>
<td>II/III</td>
<td>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.</td>
</tr>
<tr>
<td>IV</td>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>IC Methodologic Features Associated With Each Evidence Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>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 $k$; (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).</td>
</tr>
</tbody>
</table>

Acknowledgments: We thank Rachel Emery, BA; Stacie Waddell, MLS, Mary Kaye Howard, MEd, and Christine Willis, MLS, 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 Harriss, PhD, Karen Cook, PhD, Katie Snappin, MS, Grace Wang, MS, Katherine Schomer, MA, (MSKTC Quality Review); Marcel P. J. M. Dijkers, PhD, FARM, 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 (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 (e.g., all vegetative state) OR the sample is highly skewed with nonrepresentative subjects (e.g., 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 (e.g., using correlation statistics without analyzing systematic differences between raters).

APPENDIX 2: DOC TASK FORCE RATING SCHEME FOR IC COEFFICIENTS

<table>
<thead>
<tr>
<th>IC Rating</th>
<th>Cronbach $\alpha$ (4–6 Item Scale)</th>
<th>Cronbach $\alpha$ (7+ Item Scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unacceptable</td>
<td>&lt;.60</td>
<td>&lt;.70</td>
</tr>
<tr>
<td>Acceptable</td>
<td>.60–.69</td>
<td>.70–.79</td>
</tr>
<tr>
<td>Good</td>
<td>.70–.79</td>
<td>.80–.89</td>
</tr>
<tr>
<td>Excellent</td>
<td>.80+</td>
<td>.90+</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>IRR and TRR Ratings*</th>
<th>Coefficient†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unacceptable</td>
<td>&lt;.70</td>
</tr>
<tr>
<td>Acceptable</td>
<td>.70–.79</td>
</tr>
<tr>
<td>Good</td>
<td>.80–.89</td>
</tr>
<tr>
<td>Excellent</td>
<td>.90+</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Diagnostic Study Methodologic Features</th>
<th>Prognostic Study Methodologic Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Low risk of bias. A prospective study of a broad sample suspected of having DOC using a reference standard (e.g., consensus-based diagnosis or use of a standardized scale) for ease of 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.</td>
<td>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 (e.g., PVS, very severe disability) or positive outcome (e.g., 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.</td>
</tr>
<tr>
<td>II</td>
<td>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 (e.g., consensus-based diagnosis or use of a standardized assessment scale) for ease of definition of true diagnosis or scale score. 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.</td>
<td>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 (e.g., PVS, very severe disability) or positive outcome (e.g., 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.</td>
</tr>
<tr>
<td>III</td>
<td>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 DOC scale OR independent but unblind raters administer the experimental scale and reference standard OR subsets of persons are systematically removed from study (e.g., only high and low scorers are analyzed).</td>
<td>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 (e.g., only high and low scorers are analyzed) OR ≤80% of the initial sample completes the outcome measure.</td>
</tr>
</tbody>
</table>
APPENDIX 4: AAN EVIDENCE CLASSIFICATION SCHEME FOR RATING DOC DIAGNOSTIC AND PROGNOSTIC METHODOLOGIES (Cont’d)

<table>
<thead>
<tr>
<th>Class</th>
<th>Diagnostic Study Methodologic Features</th>
<th>Prognostic Study Methodologic Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>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.</td>
<td>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.</td>
</tr>
</tbody>
</table>

References

Arch Phys Med Rehabil Vol 91, December 2010


77. Pape TLB, Heinemann AW, Kelly JP, Hurder AG, Lundgren S. A measure of neurobehavioral functioning after coma. Part I: