September 2012

Continuity Assessment Record and Evaluation (CARE) Item Set: Additional Provider-Type Specific Interrater Reliability Analyses

Prepared for

Judith Tobin, PT, MBA
Centers for Medicare & Medicaid Services
Center for Clinical Standards and Quality
Mail Stop S3-02-01
7500 Security Boulevard
Baltimore, MD 21244-1850

Prepared by

Laura Smith, PhD
Anne Deutsch, RN, PhD, CRRN
Laura Hand, BS
Audrey Etlinger, MPP
Jessica Ross, MPH
Judith Hazard Abbate, PhD
Zachariah Gage-Croll, BA
Daniel Barch, MS
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709

Barbara Gage, PhD
Brookings Institution

RTI Project Number 0209853.004.002.008

Continuity Assessment Record and Evaluation (CARE) Item Set: Additional Interrater Provider-Type Specific Reliability Analyses

by Laura Smith, PhD
Barbara Gage, PhD
Anne Deutsch, RN, PhD, CRRN
Laura Hand, BS
Audrey Etlinger, MPP
Jessica Ross, MPH
Judith Hazard Abbate, PhD
Zachariah Gage-Croll, BA
Daniel Barch, MS

Federal Project Officer: Judith Tobin

RTI International

CMS Contract No. HHSM-500-2005-00291

September 2012

This project was funded by the Centers for Medicare & Medicaid Services under contract no. 500-00-1234. The statements contained in this report are solely those of the authors and do not necessarily reflect the views or policies of the Centers for Medicare & Medicaid Services. RTI assumes responsibility for the accuracy and completeness of the information contained in this report.

RTI International is a trade name of Research Triangle Institute.

ACKNOWLEDGMENTS

RTI would like to acknowledge the contributions of several organizations and individuals, without whom this work would not have been possible:

IRR/Video Pilot Participants:

Baystate Visiting Nurse Association & Hospice

Heritage Hall East NSG & Rehab

Shaughnessy Kaplan Rehabilitation Hospital

RML Specialty Hospital

Spaulding Rehabilitation Hospital

Golden Living Center

Amedisys—Tender Loving Care

Odd Fellow and Rebekah Rehabilitation Center

Thank you to all the providers who contributed to the interrater reliability data collection efforts.

Thank you also to our many colleagues across the RTI team. These colleagues include:

Margaret Stineman, University of Pennsylvania

Carole Schwartz, Rehabilitation Institute of Chicago

Sarah Couch, Ann-Marie Kamuf, and Chris Murtaugh, Visiting Nurse Service of New York

[This page intentionally left blank.]

CONTENTS

Executive Summary 1

1 Introduction and Methods 7
- 1.1 Background 7
- 1.2 Purpose 7
- 1.3 Methods 8
  - 1.3.1 Sample Selection 8
  - 1.3.2 Data Collection Protocol 9
  - 1.3.3 Item Selection for Testing 10
  - 1.3.4 Analyses 10

2 Results 13
- 2.1 Section II. Admission Information: Prior Functioning and History of Falls 13
- 2.2 Section III. Current Medical Information: Skin Integrity 15
- 2.3 Section IV. Cognitive Status, Mood, and Pain 21
- 2.4 Section V. Impairments 33
- 2.5 Section VI. Functional Status 44
  - 2.5.1 Core Function Items 44
  - 2.5.2 Supplemental Function Items 50
- 2.6 Section VII. Overall Plan of Care/Advance Care Directives 57

3 Summary and Conclusions 61

References 63

List of Tables

1-1 IRR testing providers by type/level of care 9

2-1 IRR testing: Prior functioning items and history of falls, IRR sample (CARE Tool Section 2) 14
2-2 IRR testing: Pressure ulcers at PAC admission and acute discharge, IRR sample, by provider type 18
2-3 IRR testing: Major wounds and turning surfaces at PAC admission and acute discharge, IRR sample, by provider type 20
2-4 IRR testing: Cognitive status, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4) 25
2-5 IRR testing: CAM, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4) 29
2-6 IRR testing: Mood and pain, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4) 30
2-7 IRR testing: Bowel and bladder impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5) 37
2-8 IRR testing: Swallowing impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5) 39
2-9 IRR testing: Other impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5) 41
2-10 IRR testing: Core self-care and mobility at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6) 47
2-11 IRR testing: Function—supplemental self-care, mobility, and IADLs at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6) 53
2-12 IRR testing: Patient overall status at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 7) 59

EXECUTIVE SUMMARY

E.1 Objective

The objective of this report is to describe the interrater reliability of the standardized patient assessment items in the Continuity Assessment Record and Evaluation (CARE) Item Set and to determine if there are differences in CARE item reliability by provider type, across acute care hospital and post-acute care settings: skilled nursing facilities (SNFs), inpatient rehabilitation facilities (IRFs), home health agencies (HHAs), and long-term care hospitals (LTCHs). This report includes provider-type specific analyses for all CARE items tested in the original interrater reliability analyses, adding to the previous work that analyzed interrater reliability within provider type for a select group of items only. Results from the prior analyses are included in Volume 2 of *The Development and Testing of the Continuity Assessment Record and Evaluation (CARE) Item Set: Final Report on Reliability Testing* (Gage et al., 2012). These analyses are in line with the underlying goal of the CARE assessment design, which was to develop an assessment that captured patient clinical, functional, cognitive, and social support characteristics with equal or better reliability than comparable items on the assessment instruments currently mandated in post-acute settings. Results from the current analysis and those profiled in the prior report show that, in general, the CARE Item Set performs as well as or better than existing patient assessment methods with regard to reliability and could be implemented as a standardized patient assessment form in all the patient care settings tested.

E.2 Methods

E.2.1 Design: In-person patient assessment by pairs of raters matched on discipline.

E.2.2 Setting: We worked with 27 providers in nine different geographic market areas, selected from the population of providers participating in CARE assessment data collection for the Post-Acute Care Payment Reform Demonstration (PAC-PRD).

E.2.3 Main Outcome Measures: For items with continuous responses, the outcome measure was the Pearson correlation coefficient. For categorical items, RTI International calculated kappa statistics, which indicate the level of agreement between raters using ordinal data, taking into account the role of chance agreement. The range commonly used to judge reliability based on kappa is as follows: 0 poor, 0.01–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1 almost perfect. For all items with categorical responses, unweighted kappa scores were calculated. For items with more than two response options available, weighted kappa scores were calculated using Fleiss-Cohen quadratic weights. Additionally, for items where multiple valid non-response options were available (e.g., "unknown," "not applicable," etc.), two sets of kappa statistics were calculated and reported. Because provider training did not emphasize the differences between the reasons for non-response, an additional set of kappas was calculated for these items with the non-responses excluded, to show the level of agreement observed for patients where the item was applicable.

E.3 Results

E.3.1 Section II. Admission Information: Prior Functioning and History of Falls: Items exhibited substantial to almost perfect interrater agreement overall, and by provider type.

We also found similar or higher kappas with the CARE tool than were found in prior studies of Medicare-mandated assessment tools.

- Overall, items showed substantial interrater agreement, with kappas ranging from 0.69 for unweighted kappas to 0.86 for weighted kappas.

- Kappas calculated by provider type were consistently substantial or better for providers except IRFs. IRFs consistently had lower kappas than the other provider types across these items; however, they still showed moderate agreement. In general, acute care hospitals and SNFs had the highest kappas for these measures.

- When looking at the History of Falls item in this section, the CARE tool exhibited near-perfect reliability with kappas of 0.927 to 0.948 for SNFs, whereas previous studies of items capturing similar concepts on the Minimum Data Set (MDS) 2.0 and MDS 3.0 had kappas ranging from 0 (Abt, 2003) to 0.967 (RAND Health, 2008) in the SNF setting. These items have comparable reliability to similar items on the MDS.

E.3.2 Section III. Current Medical Information: Skin Integrity: All overall kappas for the pressure ulcer and major wounds items displayed substantial agreement.

- Kappa scores for each item overall (calculated across all settings) showed moderate to almost perfect reliability across weighted and unweighted kappas where items had reportable results: 0.558 to 0.852.

- The pressure ulcer risk item (III.G1) showed substantial agreement across all setting types when observing the weighted kappa values (0.619–0.752). Unweighted kappas were lower, but likely explained by the availability of two "yes" options; one indicating risk was identified by clinical judgment and the other using formal assessment.

- All provider types had substantial kappas, at 0.70 or greater for the identification of whether a patient had one or more unhealed pressure ulcer at stage 2 or higher or unstageable (item III.G2), except for IRFs, which had moderate agreement (0.583).

- All kappas for pressure ulcer counts by stage, except for unstageable, were substantial or better overall and for LTCHs, which was the one setting with sufficient patients with pressure ulcers for kappa to be calculated.

- Major wounds showed substantial to almost perfect reliability by provider type. Results for items evaluating turning surfaces were more mixed, though the item indicating whether all turning surfaces were intact or not had moderate to almost perfect agreement.

E.3.3 Section IV. Cognitive Status, Mood, and Pain: For the majority of the items in this section, the overall kappas, both weighted and unweighted, were substantial or higher. The selected items with lower kappas are likely explained by small number of patients with the factor

being measured in the item, such as IV.A1, which asked whether the patient was comatose. The observational assessments items in this section had very low sample sizes, which precluded the calculation of kappas.

- The vast majority of cognitive status items had substantial to almost perfect reliability when evaluated overall and by provider type.

- Almost all items in the Brief Interview for Mental Status (BIMS), which evaluates cognitive status, had substantial or higher agreement (kappa greater than or equal to 0.60) overall and in all provider types, with several items with almost perfect agreement across settings. The exceptions were, for SNFs, the item indicating whether the interview was attempted, and for SNFs and IRFs, the temporal orientation item.

- Results were mixed for the Confusion Assessment Method (CAM) items, the inattention and disorganized thinking had substantial or higher agreement overall, however results were not consistent when looking at reliability within specific provider type settings. All items performed well in IRF, except the item indicating the presence of psychomotor retardation, kappa statistics showed lower agreement across items in other settings. However, the proportion of the sample that these items applied to was small in some settings, which likely impacted the observed agreement, and the ability to report results for acute and LTCH settings.

- With few exceptions when considered overall and by provider type, the mood items had kappas above 0.70, indicating substantial agreement, and many had kappas indicating almost perfect reliability (kappa greater than 0.80). SNFs had lower kappas though were 0.60 or higher, indicating at least substantial agreement.

- Pain interview items also performed well with substantial to almost perfect reliability statistics ranging over 0.70 for all items and settings except for SNFs which had slightly lower kappas, falling between 0.6 and 0.7.

- Pain observational assessment items had similar results with substantial to perfect agreement for all provider types except SNFs.

E.3.4 Section V. Impairments: Overall, many of the items in this section show very good reliability. In areas where the agreement is low, there tended to be issues of small sample size or inclusion of "not applicable" responses. Lower reliability was seen in cases where the item is not usually measured in that setting—for example, grip strength is not often measured in LTCHs. With the exception of questions related to device management, the bowel and bladder items showed substantial consistency between raters and across provider types.

- Bowel and Bladder: Kappa scores for each item when evaluated overall ranged from 0.626 to 0.896, indicating substantial to almost perfect agreement. Agreement was substantial for most provider types and items, with a few exceptions. The two items identifying need for assistance in managing bowel and bladder equipment had lower provider type specific kappas. Unweighted kappas were in the moderate range for

frequency of bladder incontinence measured in acute hospitals and HHAs. Frequency of bowel incontinence measured in IRFs and acute hospitals were lower when valid non-responses like "unknown" were kept in the analysis. Items evaluating prior use of bowel and bladder devices also had lower agreement in acute and IRFs.

- Swallowing: Results for swallowing were mixed. Agreement was perfect for the item indicating intake not by mouth (V.B1e) and the item indicated whether there were any signs or symptoms (V.B1g) had substantial to almost perfect kappas. However, agreement rates were lower and more variable across provider types for the other signs and symptoms items. SNFs had moderate to substantial kappas (0.528–0.714) for these items, and LTCHs had perfect agreement for two of the four items. Kappas were more variable for the other settings.

- Hearing, Vision, and Communication: Kappa scores for each item overall (averaged across all settings) ranged from 0.661 to 0.847, indicating substantial to almost perfect reliability. The vast majority of items had provider-type specific kappas above 0.6. The lowest kappa still indicates moderate reliability, at 0.470. In general, LTCHs and acute hospitals had the highest reliability.

- Weight Bearing: Kappa scores for each item overall (evaluated across all settings) ranged from 0.712 to 0.900, indicating substantial to almost perfect reliability. IRF and SNFs had substantial to perfect agreement on the items asking which extremity had restrictions, and were the only provider types with sufficient sample size to evaluate these items.

- Grip Strength: Kappa scores for each item overall (averaged across all settings) ranged from 0.727 to 0.885. Only LTCHs had any items with kappas scoring less than 0.6. Excluding LTCHs, the range was 0.625 to 1.00; including LTCHs, the range was 0.430 to 1.00.

- Respiratory: Kappa scores for each item overall (averaged across all settings) ranged from 0.696 to 0.874. The majority of provider types had moderate unweighted kappas but when the "not applicable" and "unknown" responses were excluded from the sample, provider types with sufficient sample size to calculate kappa had substantial to almost perfect agreement.

- Endurance: Kappa scores for each item overall (averaged across all settings) ranged from 0.539 to 0.665. Kappas were consistently substantial to almost perfect for acute and SNF settings, but consistently lower for IRFs and LTCHs.

E.3.5 Section VI. Functional Status: Functional status is divided into four sections in the CARE tool: core self-care, core motor, supplemental self-care, and supplemental motor. Items have six response levels indicating level of independence, plus five potential valid non-responses (e.g., not attempted due to medical condition).

- Over half of the core items had kappas above 0.6 (moderate reliability) when considered overall, looking at both weighted and unweighted kappas and calculated

with and without the valid non-responses. Weighted kappas were 0.700 and higher for over two thirds of the items.

- For many the core items, IRF and LTCH facilities had lower unweighted kappas than the other provider types; however, most of these differences were reduced or eliminated when responses indicating activities were not attempted were removed from the analyses. The majority of all provider type specific weighted kappas were 0.7 or better for these items.

- Most of the walk and wheel items could not be evaluated at the provider type level due to small sample sizes.

- Results were somewhat mixed for the supplemental items. Looking at the weighted kappas, excluding the valid non-responses such as "not applicable," only two items had kappas lower than 0.60 for any setting where the sample size was sufficient to report a value (LTCHs for "roll left and right" and "sit to lying"). Unweighted results were more mixed. Selected mobility and several of the instrumental activities of daily living (IADL) items had insufficient sample size because these are activities that could not be evaluated for sicker patients. Kappa scores for each item overall (averaged across all settings) ranged from 0.220 to 0.819, with most overall items in the 0.4 to 0.6 range (moderate reliability).

E.4 Summary and Conclusions

Most CARE items showed at least substantial interrater reliability, with kappas ranging from 0.6 to 0.8 for the majority of items when considered overall and by provider type. The function and impairment items showed slightly lower reliability, and for some items, reliability varied by setting. Where items showed poorer reliability in selected provider types, frequently the items applied only to a small group of patients in that setting, resulting in small sample sizes for the kappa calculation, or the condition or characteristic being measured was infrequent in that setting (e.g., comatose). Overall, however, the reliability results show that the CARE tool can be implemented in all of the settings studied, yielding patient assessment results that are at least as reliable as those produced by the previously used or currently mandated patient assessments.

The majority of items in the prior functioning; history of falls; skin integrity; and cognitive status, mood, and pain sections all showed at least substantial agreement, with kappas of between 0.6 and 0.8 on most items. Selected items that showed poor reliability in these sections frequently had small sample sizes or only a few patients had the characteristic that was being measured, which likely contributed to the low kappas.

Impairments and functional status showed substantial reliability for many items. There were more items in these sections with kappas the range of 0.4–0.6 than the other sections, however. Again, impairment and function items with lower reliability frequently had lower sample sizes than items with higher kappas. Items in the impairment and functional status sections also had more variation by provider type than other CARE items, though it was not systematic which provider types tended to have poorer reliability results, except for a few groupings of items. Consistently lower reliability was observed for IRFs and LTCHs than other

provider types for the endurance items. IRFs showed lower reliability than other provider types for the hearing, vision, or communication impairment items. For the function items, IRFs tended to show lower reliability results than the other provider types, though generally agreement rates were still moderate to substantial on these items. It is possible that the greater variation in the functional ability of the patients found in IRFs may be a cause. Agreement among raters may be easier to obtain when patients are at the highest or lowest level of dependence. In the other PAC settings, patients are more likely to be clustered at the least dependent (HHAs) and at the most dependent (LTCHs) ends of the functional scales. However, LTCHs showed low kappas for the core self-care items and moderate (but lower than the other settings) kappas for the core mobility items, which may also be partially attributable to small sample sizes. Many of the core mobility items testing the ability to walk or wheel had a sample size too small to calculate kappas. IRFs and LTCHs had the lowest reliability of provider types for the supplemental function items, and SNFs had the highest reliability among participating provider types, mostly in the 0.7–0.8 range.

Items showed a range of reliability in acute care hospitals with items in some sections indicating high reliability (Prior Function) and others having lower than average performance compared with other setting types (Bowel and Bladder Impairments). Many of the supplemental items related to instrumental activities of daily living (IADLs) (such as "laundry" or "make light meal") were not attempted or not applicable to LTCH patients, which reduced the sample size for these items. Low sample sizes in combination with low variability in responses where the items were completed (these patients were frequently dependent) likely explains the lower kappas observed for function items assessed in LTCHs.

However, overall, the reliability results show that the CARE tool can be implemented in all of the settings studied and patient assessment can be conducted with results at least as reliable as provided by the previously used patient assessment tools.

SECTION 1
INTRODUCTION AND METHODS

1.1 Background

The Centers for Medicare & Medicaid Services (CMS) has undertaken a major initiative to evaluate and realign the incentives for inpatient and post-acute services provided under the Medicare program. Currently, about a fourth of all beneficiaries are admitted to a general acute hospital each year; almost 35 percent of them are discharged to additional care in a long-term care hospital (LTCH), inpatient rehabilitation facility (IRF), skilled nursing facility (SNF), or home with additional services provided by a home health agency (HHA) (Gage et al., 2008). Although these services constitute a continuum of care for the patient, the current measurement systems do not allow Medicare to examine the effects of these continuing services on the patient's overall health and functional status.

The Deficit Reduction Act of 2005 directed CMS to address this issue and develop methods for measuring Medicare beneficiaries' health status in a consistent way that would allow CMS to examine whether Medicare's various payment systems introduced inconsistent incentives for treating clinically-similar patients. The Continuity Assessment Record and Evaluation (CARE) was developed with a standardized set of items for measuring medical, functional, cognitive, and social support factors in the acute hospital, LTCH, IRF, SNF, and HHA. These items are based on the science underlying the currently mandated assessment items in the Medicare payment systems, including those in the mandated IRF-Patient Assessment Instrument (PAI), Minimum Data Set (MDS), and Outcome and Assessment Information Set (OASIS) instruments. Additionally, the development of the CARE was based on input collected through various stakeholder meetings, including several open-door forums (ODFs) and technical expert panels (TEPs) and public comments. The CARE items were revised following a pilot test and the resulting changes were implemented for use in the Post-Acute Care Payment Reform Demonstration (PAC-PRD). Over 40,000 assessments were collected in acute hospitals, LTCHs, IRFs, SNFs, and HHAs. An additional 456 assessments were collected as part of a test of item reliability.

1.2 Purpose

An assessment tool should be both valid and reliable. It is important that items measure the concepts they were designed to capture (validity), but also that they obtain consistent results when used by different raters (reliability). Two types of reliability tests of the CARE assessment were conducted: a traditional interrater reliability test, which examines how well the items measure the specific concepts when two clinicians are measuring the same patient at the same time; and the second approach, reported on previously (Gage et al., 2012), which used videos of "standard patients" to allow examination of how discipline and setting affected item scoring. This earlier report also included results in aggregate for the traditional interrater reliability testing, with a small selection of items with reliability calculated by provider type. The current report builds on that work, by adding, for all CARE items in the original reliability report, analyses that look at how well items performed within the different post-acute settings, comparing the reliability results for acute hospitals, LTCHs, IRFs, SNFs, and HHAs. The results are important for understanding how well the standardized items perform relative to those

already used in the respective care settings to monitor the quality of care and adjust payment policies for differences in patient severity or case-mix characteristics.

1.3 Methods

As described in the initial reliability testing report (Gage et al., 2012), RTI convened a reliability working group including clinical experts in the development of existing CMS assessments and the CARE tool items (D. Saliba, A. Jette, M. Stineman, C. Murtaugh, A. Deutsch, and T. Mallinson) to help develop methods for both interrater reliability and video testing. Second, RTI conducted an extensive literature review to identify reliability standards achieved for similar items in the IRF-PAI, MDS 2.0, MDS 3.0, OASIS-B, and OASIS-C. The goal for the CARE tool results was to meet or exceed these benchmarks or past reliability levels.

1.3.1 Sample Selection

RTI estimated the required sample size for this work and determined that approximately six to eight unique providers should be recruited from each of the five levels of care (acute hospitals, HHAs, IRFs, LTCHs, and SNFs). Each provider involved in reliability testing completed a duplicate CARE tool on 15–20 PAC-PRD patients (10–15 patients in the home health setting), in accordance with the guidelines and protocols developed by RTI.

The PAC-PRD team recommended a subset of the nearly 150 providers within the PAC-PRD 12 market areas to target for reliability testing, focusing particularly on providers that were midway through their CARE data collection. RTI began actively recruiting these participating providers for CARE tool reliability testing in February 2009. Nine of the 12 market areas were included in the reliability sample, allowing for efficiencies and ensuring that the included providers were geographically diverse. RTI recruited 27 providers from the set of providers already enrolled in the PAC-PRD data collection. See ***Table 1-1*** for counts of providers and the number of assessment pairs submitted by each provider type. The number of participants of each type reflected participation levels in the PAC-PRD data collection and were consistent with reliability sample sizes in the benchmark studies. Facilities were asked to enroll a set number of fee-for-service (FFS) Medicare patients each month, representing a range of function and acuity. Providers with low Medicare admissions or that had only one clinician conducting CARE assessments (and therefore would not be able to conduct a paired assessment with another clinician) were not recruited.

Table 1-1
Interrater reliability testing providers by type/level of care

Provider type	Number of providers enrolled	Paired assessment numbers
Acute hospitals	4	66
Home health agencies (HHAs)	8	102
Inpatient rehabilitation facilities (IRFs)	7	118
Long-term care hospitals (LTCHs)	2	49
Skilled nursing facilities (SNFs)	6	121
Total	**27**	**456**

1.3.2 Data Collection Protocol

RTI considered two approaches to examine the interrater reliability of CARE items: a "gold standard" methodology and a within-setting paired rater methodology. The use of "gold standard" data collectors is a common approach. Under this method, a small number of clinicians, usually nurses, are provided intensive training on the instrument and the interrater reliability of these raters is examined and retraining provided until they are quite consistent with each other. These "gold standard" raters are then sent to facilities where they observe and score patients and their ratings are compared with those of the facility nurses. The strength of this approach, comparison to a "gold standard" rater, is also its weakness. Because these "gold standard" raters undergo very expensive and extensive training to achieve their high level of rating consistency and accuracy, data collected by clinicians in the field, who generally have not had this level of training, will fall short of this level of accuracy. Yet it is these data from the field that will be the basis of both the demonstration sample that will develop the payment models, and the data that will subsequently be submitted to CMS for reimbursement. These data reflect the "practicably achievable" level of reliability, rather than an idealized standard.

RTI therefore used a traditional interrater reliability method that compares pairs of raters within each site. Under this method, two raters observe the same patient, or review the same chart, then independently assign ratings. Interrater reliability was tested for differences across the acute and post-acute care settings, to determine if there were significant differences in the reliability of the items in an IRF versus an SNF, HHA, or LTCH. The strength of this approach is that the ratings reflect standards and performance of clinicians in the field. The challenge of this approach is that it is costly in terms of staff time because two clinicians must be available to observe each patient for a given time period. Providers were instructed to have pairs of raters complete both patient assessments at the same time upon admission or, at a minimum, within the 48-hour reference data window. Only staff previously collecting CARE information in the demonstration participated in interrater reliability testing. Each demonstration site identified two or three clinicians in each setting, and each clinician was primary observer on five cases and secondary observer on another five cases. Patients were assessed by staff pairs matched by discipline (two nurses, two physical therapists, etc.). To account for different lengths of time

elapsed since the initial PAC Demonstration CARE training in each market, each clinician participating in interrater reliability testing attended a 1.5-hour CARE refresher training prior to beginning the data collection. Following CARE refresher training, RTI also reviewed the data collection protocol with the demonstration project coordinators.

Responses to items in the CARE tool were obtained by one or more of the following predetermined, matched methods: direct observation of the patient (includes hands-on assistance), patient interviews (with each team member taking turns conducting and observing patient interviews), interviews with relatives/caregiver of the patient for certain items, and interviews with staff caring for the patient and/or chart review. Rater pairs were instructed to determine in advance which methods would be used to score the particular CARE tool items and to have both raters use the same methods. Raters were encouraged to divide hands-on assistance to the patient as evenly as possible for CARE items that required hands-on assistance, such as the functional status item "Sit to Stand." For patient interview items, such as those in the temporal orientation/mental status, mood, and pain sections, raters were instructed that one rater could conduct the entire interview, or the raters could alternate questioning. Note that by having the raters rate the patient simultaneously, only one of the clinicians is performing the assessment as it is meant to be performed (e.g., reading the questions to the patient, or providing assistance to the patient), whereas the other clinician is listening to the questions being read. This discrepancy between the test conditions and the intended use may result in differences in the performance of the measures from regular use in the field. This may be less of a concern for the function items because they are intended to be rated based on the patient's *usual* (but still observed) performance, which would not necessarily correspond to the performance observed at the moment of assessment. Raters were instructed not to discuss CARE item scoring during the CARE assessment, nor to share item scores until the data were entered into the CMS database and finalized. Providers submitted CARE data via the online CARE application for both assessments in each pair and submitted a list of assessment IDs associated with both the PAC-PRD assessment and the paired interrater reliability assessment on paper.

RTI initially conducted a small pilot in the Boston market area to test and refine the protocol, refresher training, and checklists.

1.3.3 Item Selection for Testing

CARE tool items selected for interrater reliability testing fell into one (or more) of the following categories: items that are subjective in nature, items that have not previously appeared in CMS tools (i.e., new CARE items), items that influence payments or are used in payment models currently, or items not previously tested in certain settings.

For the duplicate assessment, raters from HHAs, SNFs, IRFs, and LTCHs completed a CARE Tool Admission Form on each patient enrolled in reliability testing. Raters from acute hospitals used an Acute Care Discharge Form.

1.3.4 Analyses

RTI used two analytic approaches for assessing the interrater reliability of the CARE tool items, following closely the methods used in prior CMS assessment interrater reliability analyses. For continuous items, RTI calculated Pearson correlation coefficients to show the

extent of agreement between two raters on the same item. For categorical items, RTI calculated kappa statistics, which indicate the level of agreement between raters using ordinal data, taking into account the role of chance agreement. The range commonly used to judge reliability based on kappa is as follows: 0 poor, 0.01–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1 almost perfect.

For categorical items with only two responses available, RTI calculated unweighted kappas. For items with more than two responses, RTI calculated both weighted and unweighted kappas. Unweighted kappa assumes the same "distance" between every one unit difference in response across an ordinal scale (e.g., for the CARE functional item scale range 1–6, an unweighted kappa assumes the difference in functional ability between a score of 1 = dependent and 2 = substantial/maximal assistance is the same as the difference in functional ability between 5 = setup or cleanup assistance and 6 = independent). RTI used Fleiss-Cohen weights, or quadratic weights, which approximate the intra-class correlation coefficient and are commonly used for calculating weighted kappas. This choice of weighting is consistent with prior analyses of assessment reliability where the method for developing weights was specified (see Hirdes et al., 2002, and Streiner and Norman, 1995). Note that Fleiss-Cohen weights put lower emphasis on disagreements between responses that fall "near" to each other on an item scale. It should also be noted that the value of kappa can be influenced by the prevalence of the outcome or characteristic being measured. If the outcome or characteristic is rare, the kappa will be low because kappa attributes the majority of agreement among raters to chance. Kappa is also influenced by bias, and if the effective sample size is small, variation may also play a role in the results. Hence, we report both weighted and unweighted kappas to show the range of agreement found under the two sets of assumptions.

Additionally, RTI calculated a separate set of kappa statistics (unweighted and weighted where applicable) for items where additional responses outside of an ordinal scale were available (letter codes) and were set to missing. For example, for Section 6, Functional Status items of the CARE tool, providers could choose between five and six different letter codes designating that an item was "not attempted." Because training did not emphasize distinctions between these letter code responses and these responses were not necessarily ordered, we are reporting a set of kappas for these items where the "not attempted" responses are recoded to missing. The results of these analyses are described in the following section. We also compare the CARE results with interrater reliability testing available from prior studies for similar items on Medicare mandated assessments. (Also see Appendix A of Volume 2 of *The Development and Testing of the Continuity Assessment Record and Evaluation [CARE] Item Set: Final Report on Reliability Testing* for additional detail.)

[This page intentionally left blank.]

SECTION 2
RESULTS

2.1 Section II. Admission Information: Prior Functioning and History of Falls

Capturing patients' functional status prior to admission is relevant for understanding patient outcomes, particularly the patient's potential for functional improvement during a treatment period. Prior function measures in the Continuity Assessment Record and Evaluation (CARE) tool include the ability to perform everyday activities such as self-care, mobility (ambulation and wheelchair), stairs, and functional cognition prior to the current illness, exacerbation, or injury. ***Table 2-1*** shows the results for the prior functioning items and history of falls, both overall and across provider types (II.B5 and II.B7). Two sets of data are presented—the first three columns present the data including the cases where the response code was "not applicable"; the second set of columns present the kappas with only the rated cases (excluding the "N/A" cases).

The self-care item (II.B5a) displayed simple and weighted kappas for acute care hospitals and skilled nursing facilities (SNFs) that indicate almost perfect agreement, and long-term care hospitals (LTCHs) and home health agencies (HHAs) had substantial agreement in analyses with and without the "unknown" and "not applicable" responses included. Kappas for inpatient rehabilitation facilities (IRFs) show a lower reliability but moderate level of agreement for self-care. The Mobility and Stairs (Ambulation) items (II.B5b and II.B5c) showed almost perfect agreement for acute care hospitals, LTCHs, and SNFs when the "not applicable" and "unknown" responses were excluded. Simple and weighted kappas for HHAs show substantial agreement, and IRFs again display a moderate, but still positive, agreement for these items.

For the wheelchair mobility item (II.B5d), LTCHs and SNFs show almost perfect agreement in the weighted kappa statistics (0.922 and 0.910, respectively) and perfect agreement when the "unknown" and "not applicable" responses are removed. The simple and weighted kappas for acute hospitals also show almost perfect agreement (the weighted kappa is 0.970), whereas those for HHAs show substantial agreement. The functional cognition item (II.B5e) includes almost perfect weighted kappas for SNFs and HHAs. Acute hospitals and LTCHs show substantial agreement. Finally, for item II.B7, History of Falls, the simple and weighted kappas for acute hospitals, SNFs, and HHAs show almost perfect agreement, and LTCHs display substantial agreement. In all of these cases, IRFs demonstrate slightly lower but still moderate levels of agreement.

Summary

In this section, we discussed the overall and provider-type specific kappas for items relating to prior functioning and history of falls.

- The overall items exhibited substantial interrater agreement, with kappas ranging from 0.69 in unweighted kappas to 0.86 in weighted kappas.

- Kappas were fairly consistent within each of the five types of providers. Though IRFs consistently had lower kappas than the other provider types, they still showed

moderate agreement. In general, acute care hospitals and SNFs had the highest kappas for these measures.

- When looking at the History of Falls item in this section, we found similar or higher kappas with the CARE tool than were found in prior studies of similar items on the MDS 2.0 and MDS 3.0. The CARE tool exhibited near-perfect reliability with kappas of 0.927 to 0.948 for SNFs, whereas previous studies of items capturing similar concepts on the MDS 2.0 and MDS 3.0 had kappas ranging from 0 (Abt, 2003) to 0.967 (RAND Health, 2008) in the SNF setting.

Table 2-1
IRR testing: Prior functioning items and history of falls, IRR sample
(CARE Tool Section 2)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
Prior Functioning						
II.B5a Self-care	442	0.749	0.761	427	0.773	0.795
Acute	60	0.917	0.887	60	0.917	0.887
HHA	100	0.685	0.733	99	0.699	0.737
IRF	115	0.484	0.432	111	0.536	0.502
LTCH	49	0.758	0.799	42	0.821	0.785
SNF	118	0.900	0.860	115	0.910	0.943
II.B5b Mobility (ambulation)	442	0.731	0.696	412	0.729	0.752
Acute	60	0.772	0.428	55	0.856	0.895
HHA	100	0.616	0.610	96	0.586	0.597
IRF	115	0.626	0.570	105	0.559	0.414
LTCH	49	0.809	0.833	45	0.810	0.802
SNF	118	0.862	0.811	111	0.903	0.949
II.B5c Stairs (ambulation)	442	0.719	0.739	292	0.781	0.863
Acute	60	0.885	0.765	40	1.000	1.000
HHA	100	0.690	0.680	74	0.750	0.861
IRF	115	0.426	0.446	73	0.509	0.525
LTCH	49	0.746	0.825	28	0.848	0.938
SNF	118	0.906	0.937	77	0.899	0.937
II.B5d Mobility (wheelchair)	441	0.693	0.807	86	0.823	0.845
Acute	60	0.836	0.970	†	†	N/A
HHA	100	0.699	0.727	17	0.788	0.837
IRF	114	0.365	0.462	†	†	N/A
LTCH	49	0.743	0.922	16	1.000	1.000
SNF	118	0.871	0.910	30	1.000	1.000

(continued)

Table 2-1 (continued)
IRR testing: Prior functioning items and history of falls, IRR sample
(CARE Tool Section 2)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
II.B5e Functional cognition	441	0.701	0.737	413	0.746	0.803
Acute	60	0.664	0.621	58	0.705	0.760
HHA	100	0.676	0.808	96	0.725	0.852
IRF	114	0.454	0.491	107	0.515	0.522
LTCH	49	0.615	0.615	38	0.627	0.441
SNF	118	0.918	0.827	114	0.927	0.928
Falls						
II.B7 History of falls	431	0.839	0.764	402	0.876	N/A
Acute	60	0.935	0.941	59	0.932	N/A
HHA	95	0.874	N/A	95	0.874	N/A
IRF	112	0.689	0.525	101	0.773	N/A
LTCH	49	0.739	0.651	37	0.853	N/A
SNF	115	0.936	0.948	110	0.927	N/A

NOTE: N/A—Weighted kappa is not applicable for items with only two responses available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; SNF, skilled nursing facility.

**With unknown and not applicable responses excluded.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data, IRR sample only (CARE extract 1/28/10).

2.2 Section III. Current Medical Information: Skin Integrity

Skin integrity issues comprise a major source of patient complications, affecting both resource needs and patient outcomes. The CARE tool includes two core items on pressure ulcers, which indicate whether the patient is at risk of developing pressure ulcers and whether they have one or more unhealed pressure ulcers at stage 2 or higher. The supplemental items include the proportion of patients with pressure ulcers who had stage 2, 3, or 4 ulcers. The pressure ulcer items were developed with input from representatives from the Wound, Ostomy, and Continence Nurses (WOCN) and the National Pressure Ulcer Advisory Panel (NPUAP). The tool also includes a core item assessing the presence of major wounds and supplemental items designed to further characterize the types of major wounds that may be present. Supplemental items are only reported for cases having a core item present. For example, the supplemental items indicating presence of any diabetic foot ulcers or vascular ulcers reflect the severity of wound issues within the population who had at least one major wound.

Results for this section are displayed in ***Tables 2-2*** and ***2-3***. Note that the correlations are reported for "Longest length of the largest stage 3 or 4 pressure ulcer" (III.G3a) and "Longest width of the largest stage 3 or 4 pressure ulcer" (III.G3b), rather than kappas, because these are

continuous variables, not categorical items. All overall kappas for the pressure ulcer items evaluated indicate substantial or near-perfect consistency except for item indicating any stage 3 or 4 pressure ulcer with undermining and/or tunneling (III.G4), which has fewer than 11 cases, so the kappa was not calculated. Correlations for the length and width of the most problematic pressure ulcer are, however, relatively high, showing moderate reliability at 0.596 and 0.578, respectively. Similarly, overall kappas on the Turning Surfaces (III.G6a-e) are also relatively high with moderate agreement for each item except the "Other surfaces not intact" (III.G6e). This item was less specific than the other turning surfaces item and infrequently indicated, which may have contributed to the lower agreement statistic.

Provider-type specific kappa scores were fairly consistent across settings, with most indicating substantial or higher agreement. IRFs had slightly lower kappas though were still in the moderate range. For the pressure ulcer item "Does this patient have one or more unhealed pressure ulcer(s) at stage 2 or higher or unstageable" (III.G2), kappas for HHAs, LTCHs, and SNFs each indicate almost perfect agreement. Kappas for acute hospitals demonstrate substantial agreement, whereas interrater reliability in IRFs was the lowest among the five provider types with kappas indicating moderate concurrence on the item among clinicians at each of these facilities.

There were generally few patients in the Post-Acute Care Payment Reform Demonstration (PAC-PRD) sample with pressure ulcers; therefore, the pressure ulcer items (III.G2a through III.G4) tended to have small sample sizes for most provider types and did not allow for kappa analyses. The major wound indicator (III.G5) exhibited substantial or almost perfect agreement across all provider types. For CARE tool item "Skin for all turning surfaces is intact" (III.G6a), LTCHs exhibit almost perfect consensus between raters, whereas kappas for both acute care providers and HHAs indicate substantial agreement. Kappas for IRFs and SNFs demonstrate moderate agreement between clinicians in each of these care settings. The items on turning surfaces (III.G6a through III.G6e) show slightly more variation—on a few items the kappas were unable to be interpreted. However, the item indicating whether all turning surfaces were intact or not (III.G6d) exhibited almost perfect agreement in every provider type except IRFs, which still had substantial agreement with a kappa of 0.749.

Summary

This section discussed the overall and by-provider-type reliability of the skin integrity items from the CARE tool.

- All overall kappas for the pressure ulcer and major wounds items displayed substantial to almost perfect agreement, with the exception of the item indicating the presence of a pressure with undermining and/or tunneling present (III.G4), which had a sample size too small to calculate a kappa.

- The correlation for the length of the most problematic pressure ulcer showed moderate agreement at 0.596.

- For the turning surfaces items, the overall kappas were slightly lower, but most were still in the substantial agreement range. Only two overall kappas were lower:

III.G6b, "Right hip not intact," showed moderate agreement; and III.G6e, "Other turning surface(s) not intact," showed fair agreement.

When looking at the provider-type specific results, most responses were substantial or higher, and several of the lower kappas may be explained by small sample sizes.

- The pressure ulcer risk item (III.G1) showed substantial agreement across all setting types when observing the weighted kappa values (0.619–0.752). Unweighted kappas were lower, but likely explained by the availability of two "yes" options; one indicating risk was identified by clinical judgment and the other using formal assessment.

- All provider types had substantial kappas, at 0.70 or greater for the identification of whether a patient had one or more unhealed pressure ulcer at stage 2 or higher or unstageable (III.G2), except for IRFs, which had moderate agreement (0.583).

- All kappas for pressure ulcer counts by stage, except for unstageable, were substantial or better overall and for LTCHs, which was the one setting with sufficient patients with pressure ulcers for kappa to be calculated.

- Major wounds (III.G5) had substantial to almost perfect reliability by provider type.

- Results for items evaluating turning surfaces were more mixed, though the item indicating whether all turning surfaces were intact or not had moderate to almost perfect agreement. The "Back/buttocks not intact" item (III.G6d) had the highest provider type specific kappas and the "Other turning surface(s) not intact" (III.G6e), had the lowest, with all provider-type specific kappas indicating fair agreement or lower.

- The CARE item identifying "Any stage 2+ pressure ulcers" (III.G2) showed similar, almost perfect, reliability for SNFs (kappa = 0.815) to studies of a comparable item on the MDS 2.0 by Abt Associates (2003) and Mor et al. (2003), which both yielded kappas of 0.83. The CARE item had a higher kappa than the Staff Time and Resource Intensity Verification (STRIVE) study (Iowa Foundation of Medical Care, n.d.) of the MDS 2.0 (kappa = 0.52).

Table 2-2
IRR testing: Pressure ulcers at PAC admission and acute discharge, IRR sample, by provider type

Item	Effective sample size	Kappa	Weighted kappa
Pressure Ulcers			
III.G1 Is the patient at risk of developing pressure ulcers?	450	0.586	0.742
Acute	65	0.425	0.619
HHA	101	0.573	0.681
IRF	116	0.473	0.705
LTCH	48	0.579	0.752
SNF	120	0.586	0.636
III.G2 Does this patient have one or more unhealed pressure ulcer(s) at stage 2 or higher or unstageable?	447	0.845	N/A
Acute	63	0.734	N/A
HHA	101	0.889	N/A
IRF	116	0.583	N/A
LTCH	49	0.916	N/A
SNF	118	0.815	N/A
Number of pressure ulcers present at assessment by stage			
III.G2a Stage 2	44	0.815	0.801
Acute	†	†	N/A
HHA	†	†	N/A
IRF	†	†	N/A
LTCH	19	0.627	N/A
SNF	†	†	N/A
III.G2b Stage 3	43	0.852	0.760
Acute	†	†	N/A
HHA	†	†	N/A
IRF	†	†	N/A
LTCH	18	0.613	0.348
SNF	†	†	N/A
III.G2c Stage 4	43	0.780	0.707
Acute	†	†	N/A
HHA	†	†	N/A
IRF	†	†	N/A
LTCH	18	0.700	0.634
SNF	†	†	N/A
III.G2d Unstageable	43	0.652	0.678
Acute	†	†	N/A
HHA	†	†	N/A
IRF	†	†	N/A

(continued)

Table 2-2 (continued)
IRR testing: Pressure ulcers at PAC admission and acute discharge, IRR sample, by provider type

Item	Effective sample size	Kappa	Weighted kappa
LTCH	18	0.417	0.579
SNF	†	†	N/A
III.G2e Unhealed stage 2 or higher pressure ulcers present more than 1 month	41	0.790	0.825
Acute	†	†	N/A
HHA	†	†	N/A
IRF	†	†	N/A
LTCH	17	0.558	0.609
SNF	†	†	N/A
Longest length and width of stage 3 or 4 unhealed pressure ulcer (correlations)			
III.G3a Longest length	22	0.596	N/A
Acute			
HHA	15	0.250	—
IRF	†	†	—
LTCH	—	—	—
SNF	—	—	—
III.G4 Undermining and or tunneling present	†	†	—
Acute	†	†	—
HHA	†	†	—
IRF	†	†	—
LTCH	†	†	—
SNF	†	†	—

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

† Kappas and correlations for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data retrieved from programming request LS20 [CAREREL063] and LS09 [CAREREL030]).

Table 2-3
IRR testing: Major wounds and turning surfaces at PAC admission and acute discharge, IRR sample, by provider type

Item	Effective sample size	Kappa	Weighted kappa
Major Wounds			
III.G5 One or more major wounds that require ongoing care	378	0.789	N/A
Acute	58	0.733	N/A
HHA	66	0.828	N/A
IRF	111	0.693	N/A
LTCH	43	0.790	N/A
SNF	100	0.852	N/A
Turning Surfaces Not Intact			
III.G6a Skin for all turning surfaces is intact	451	0.665	N/A
Acute	65	0.642	N/A
HHA	101	0.718	N/A
IRF	116	0.523	N/A
LTCH	49	0.876	N/A
SNF	120	0.598	N/A
III.G6b Right hip not intact	451	0.558	N/A
Acute	65	#	N/A
HHA	101	0.000[a]	N/A
IRF	116	0.525	N/A
LTCH	49	0.545	N/A
SNF	120	0.585	N/A
III.G6c Left hip not intact	451	0.630	N/A
Acute	65	#	N/A
HHA	101	0.000[b]	N/A
IRF	116	0.645	N/A
LTCH	49	0.539	N/A
SNF	120	1.000	N/A
III.G6d Back/buttocks not intact	451	0.766	N/A
Acute	65	1.000	N/A
HHA	101	0.951	N/A
IRF	116	0.749	N/A
LTCH	49	0.821	N/A
SNF	120	0.497	N/A
III.G6e Other turning surface(s) not intact	451	0.208	N/A
Acute	65	0.000	N/A
HHA	101	−0.020[c]	N/A

(continued)

Table 2-3 (continued)
IRR testing: Major wounds and turning surfaces at PAC admission and acute discharge, IRR sample, by provider type

Item	Effective sample size	Kappa	Weighted kappa
IRF	116	0.169	N/A
LTCH	49	0.295	N/A
SNF	120	0.314	N/A

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

[a] Kappa unable to be interpreted. There were 101 agreements and 0 disagreements.

[b] Kappa unable to be interpreted. There were 100 agreements and 1 disagreement.

[c] Kappa unable to be interpreted. There were 97 agreements and 4 disagreements.

No discordant pairs and no variation in responses.

SOURCE: RTI analysis of CARE data (data retrieved from programming request LS20 [CAREREL063] and LS09 [CAREREL030]).

2.3 Section IV. Cognitive Status, Mood, and Pain

Measures of mental status, including cognitive function, are an important part of clinical assessment, especially in geriatrics, neurology, and medical rehabilitation. A patient's mental status not only affects their ability to interact with the clinicians and understand treatments, but also plays an important role in their ability to self-report problems such as mood and pain.

The CARE tool features multiple items used to assess a patient's cognitive status, including an assessment of persistent vegetative state (comatose); the Brief Interview for Mental Status (BIMS); an observational assessment of cognitive status; and the Confusion Assessment Method (CAM). Among these, only the comatose item is a core item assessed on the entire CARE population. Patients able and willing to respond to interview questions are assessed using the BIMS, which evaluates the ability to repeat three words, temporal orientation, and recall. The BIMS items present in the CARE tool are based largely on those developed for the MDS 3.0, with only minor adaptations made to ensure applicability to the full range of post-acute care providers. When a patient is unable or unwilling to be assessed by the BIMS, the clinician evaluates their cognitive status using the "Observational Assessment of Cognitive Status," reporting the patient's usual ability to recall the current season, staff names and faces, the location of their own room, and so forth. In turn, the CAM was triggered only when responses to the BIMS suggest the presence of cognitive impairment. The CAM, which is also derived from a similar measure on the MDS 3.0, is used to identify symptoms of delirium and subdelirium.[1]

The mood items on the CARE tool include items from the Patient Health Questionnaire-2 (PHQ-2), a validated depression screening tool for older populations, and one item ("Feeling

[1] The CAM item was included as a core item in Phase 2 of the PAC-PRD based on feedback from the participating clinicians that this item should be assessed on all patients, not restricted to those triggered by the BIMS items.

Sad") from the National Institutes of Health Patient Reported Outcomes Measurement Information System (NIH PROMIS) initiative. Mood items are included on the CARE tool because they are predictive of resource utilization and may affect outcomes. These are asked only in the PAC populations because measuring them at the time of discharge from acute hospital was considered problematic from a quality of care standpoint. Among these items, only the item for "Mood Interview Attempted" is reported for all patients.

Table 2-4 displays the results from the Cognitive Status items in Section 4 of the CARE tool. Results are for patients who were not reported as being in a vegetative state as indicated in item IV.A1. Less than 11 patients in the interrater reliability sample were comatose, which likely explains the low overall kappa for "Persistent Vegetative State/no discernible consciousness at time of admission" (IV.A1) (0.398, fair agreement), and the low kappas for all provider types except LTCH. Kappa statistics, as stated previously, are impacted by the prevalence of the factor being measured in a sample population. In this section, the kappas for SNFs were generally lower than other provider types, but overall showed at least moderate agreement.

The overall kappas were highest for the "Temporal Orientation" items (IV.B3b) and recall of three words (IV.B3c). Because the Observational Assessment of Cognitive Status and CAM are administered only to selected patients based on their responses to the BIM-related items, the sample sizes for these items are much smaller. The Observational Assessment showed no discordant assessment pairs for patients' ability to recall the current season, location of own room, and staff names and faces. The CAM, shown in ***Table 2-5***, had substantial agreement for inattention and disorganized thinking; however, altered levels of consciousness and psychomotor retardation were lower, showing moderate reliability at 0.58 and 0.48, respectively.

For the BIMS CARE tool item "Recalls year" (IV.B3b.1/ IV.B2b1), weighted kappas indicate almost perfect agreement for all inpatient hospitals (acute care, IRFs, and LTCHs), with scores ranging from 0.91 to 1.00. Participating HHAs and SNFs each had substantial agreement for the item with unweighted kappas of 0.73 and 0.62, respectively, whereas the weighted kappa for HHAs (0.90) indicates almost perfect agreement. For the CAM item "Inattention" (IV.D1), the unweighted kappa for HHAs indicates moderate agreement whereas the weighted kappa indicates substantial agreement. The kappa for IRFs was higher in the weighted version as well, indicating almost perfect agreement as compared with the substantial agreement indicated by the unweighted kappa. Both the simple and weighted kappas for LTCHs demonstrate substantial interrater agreement. Last, SNFs' unweighted kappa indicates moderate agreement on the item whereas the weighted kappa indicates substantial agreement.

Table 2-6 shows the results from the Mood and Pain section of the CARE tool. The table includes the core item for "Mood Interview Attempted," and also displays the results for the core items on "Little interest or pleasure in doing things" and "Feeling down, depressed, or hopeless?" These two questions make up the PHQ-2. When a patient responded "Yes" to either of these questions, a subsequent supplemental question was asked concerning the frequency of these feelings (CARE items F2b and F2d). Possible answers range from "Not at all," which is coded as "0," to "Nearly every day," which is coded as "3." In addition to the PHQ-2 questions, all post-acute care patients who could be interviewed also answered the core item on "Feeling

Sad." Overall and provider-type specific kappas ranged from 0.63 to 0.94 for this set of items, showing substantial to almost perfect agreement.

For the PHQ-2 item "Feeling down, depressed, or hopeless" (IV.F2c), overall kappas with "unable to answer" or "no response" excluded indicate almost perfect agreement with values ranging from 0.81 to 0.89 for all provider types except acute hospitals, which did not have this item on their tool (weighted kappas are not applicable because there are only two possible responses for the variable with excluded answers). Analyses with "unable to answer" or "no response" categories of the variable included resulted in unweighted kappas that indicate almost perfect agreement between clinicians in HHAs, IRFs, LTCHs, and SNFs, whereas weighted kappas indicate almost perfect agreement in IRFs and SNFs and substantial agreement in LTCHs. There was no weighted kappa computed for HHAs with the "unable to answer" or "no response" categories included because respondents used only two levels of the variable. Again, this item was not applicable for acute hospitals. Analyses of the rest of the PHQ-2 questions yielded similar results, with all available provider-type specific kappas having at least substantial agreement, and many with almost perfect agreement.

The CARE included both the PHQ-2 and the PROMIS items to identify whether one was more reliable than another with these populations. The PROMIS item was based on the Short Form (SF)-36, which was developed for the general population, including the healthy as well as this population where everyone is receiving acute or post-acute care. The kappas suggest the PHQ-2 items were slightly more reliable across the range of populations than the Feeling Sad item (more kappas above 0.80 although the lowest kappa on the "Feeling Sad" item was 0.742), suggesting both are substantially reliable in these populations.

Identifying the presence and severity of pain is not only critical for understanding severity of illness and anticipating resource utilization, but is also an important quality of care domain. The CARE tool includes items measuring three domains of pain: a core item asked of all patients: "Presence of pain in last 2 days," and two supplemental items asked of patients who answered yes to the core pain item: severity of pain and effect of pain on function. If a patient indicated pain was present, they were asked to categorize that pain using a 0–10 scale. The effect of pain on sleep and activities was also assessed for these patients. Clinical observation was used to determine the possible presence of pain for patients who could not be interviewed.

The interview-based pain items (IV.G1 through IV.G5) had substantial to almost perfect kappas whether coded non-response items were included in calculations or not (weighted kappa range: 0.61–0.91). Observational assessment items had lower kappa values than the interview items, as expected, but were still substantial for non-verbal sounds, vocal complaints of pain, and facial expressions (range 0.61–0.66, substantial agreement). "Protective body movements or postures" (IV.G6d) had a lower kappa at 0.42. For item "Pain presence: Pain during the last 2 days?" (IV.G2), kappas with "unable to answer" or "no response" excluded indicate almost perfect agreement (ranging from 0.91–0.94) in all care settings except for SNFs, whose kappa value indicates substantial agreement. Simple and weighted kappas with "unable to answer" and "no response" included indicate almost perfect agreement among the clinicians in each inpatient hospital (acute hospitals, IRFs, and LTCHs) and in HHA settings, and moderate agreement in SNFs.

Summary

This section discussed the cognitive status, mood, and pain items on the CARE tool.

- Overall, agreement was very good for this section on most items across settings. Most items displayed substantial or almost perfect agreement, particularly in the BIMS items and the PHQ-2 items.

- Almost all items in the Brief Interview for Mental Status (BIMS), had substantial or higher agreement (kappa greater than or equal to 0.60) overall and in all provider types, with several items with almost perfect agreement across settings. The exceptions were, for SNFs, the item indicating whether the interview was attempted, and for SNFs and IRFs, the temporal orientation item.

- Results were mixed for the Confusion Assessment Method (CAM) items, the inattention and disorganized thinking had substantial or higher agreement overall, however results were not consistent when looking at reliability within specific provider type settings. All items performed well in IRF, except the item indicating the presence of psychomotor retardation, kappa statistics showed lower agreement across items in other settings. However, the proportion of the sample that these items applied to was small in some settings, which likely impacted the observed agreement, and the ability to report results for acute and LTCH settings.

- With few exceptions when considered overall and by provider type, the mood items had kappas above 0.70, indicating substantial agreement, and many had kappas indicating almost perfect reliability (kappa greater than 0.80). SNFs had lower kappas though were 0.60 or higher, indicating at least substantial agreement.

- Pain interview items also performed well with substantial to almost perfect reliability statistics ranging over 0.70 for all items and settings except for SNFs which had slightly lower kappas, falling between 0.6 and 0.7.

- Pain observational assessment items had similar results with substantial to perfect agreement for all provider types except SNFs.

- When comparing the CARE items to current or previously mandated assessment tools, reliability was shown to be similar or higher in the CARE items.

- Berg's (1999) analyses of the OASIS-B tool items capturing cognitive, behavioral signs and symptoms, mood, and pain domains yielded comparable or lower kappas than found for the CARE tool in HHAs.

- The CARE tool had a higher range of kappas for IRFs on the disorganized thinking on the CAM than an analogous item on FIM® tool,[2] as found in the Hamilton et al. analysis (1994).

- The CARE tool's reliability as compared with analyses of the MDS 2.0 and 3.0 is slightly less consistent. On CAM items the CARE tool's kappas were similar to some studies but lower than others. However, for behavior items and pain items, the CARE items performed similarly to the reliability found in the MDS 2.0 studies (Abt, 2003; Mor et al., 2003; Morris et al., 1997). Kappas were higher for the MDS 3.0 testing, however these differences are likely attributable differences in study design, given the similarity in items on the two assessments. The MDS 3.0 study used gold-standard nurses (RAND Health, 2008); whereas, the CARE used staff for both raters to get a better reflection of how the tool would perform during regular data collection.

Table 2-4
IRR testing: Cognitive status, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa
IV.A1 Comatose	451	0.398	N/A
Acute	65	0.000	N/A
HHA	101	0.000[a]	N/A
IRF	115	#	N/A
LTCH	49	1.000	N/A
SNF	121	0.000	N/A
BIMS			
IV.B1a Interview attempted	447	0.771	N/A
Acute	64	0.660	N/A
HHA	100	1.000	N/A
IRF	115	0.786	N/A
LTCH	48	0.921	N/A
SNF	120	0.423	N/A
IV.B1b Indicate reason that the interview was not attempted	20	0.713	0.632
Acute	†	†	†
HHA	†	†	†

(continued)

[2] FIM® is a trademark of Uniform Data System for Medical Rehabilitation, a division of UB Foundation Activities, Inc.

Table 2-4 (continued)
IRR testing: Cognitive status, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa
IRF	†	†	†
LTCH	†	†	†
SNF	†	†	†
Temporal Orientation/Mental Status			
IV.B3a Repetition of three words (sock, blue, bed)	356	0.625	0.705
Acute	—	—	—
HHA	97	0.691	0.472
IRF	106	0.552	0.594
LTCH	40	1.000	1.000
SNF	113	0.556	0.770
BIMS			
IV.B3b.1/IV.B2b1 Recalls year	419	0.820	0.876
Acute	62	0.946	0.919
HHA	98	0.739	0.902
IRF	106	0.942	0.952
LTCH	40	1.000	1.000
SNF	113	0.628	0.734
IV.B3b.2/ IV.B2b2 Recalls month	419	0.790	0.869
Acute	62	0.820	0.943
HHA	98	0.808	0.939
IRF	106	0.810	0.884
LTCH	40	1.000	—
SNF	113	0.680	0.725
IV.B3b.3 Recalls day	356	0.876	N/A
Acute	—	—	—
HHA	98	0.896	N/A
IRF	106	0.863	N/A
LTCH	40	0.939	N/A
SNF	112	0.846	N/A
Recall of Three Words (Sock, Blue, Bed)			
IV.B3c.1 Recalls “sock”	357	0.829	0.895
Acute	—	—	—
HHA	98	0.912	0.921
IRF	106	0.832	0.907
LTCH	40	0.815	0.946
SNF	113	0.759	0.825

(continued)

Table 2-4 (continued)
IRR testing: Cognitive status, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa
IV.B3c.2 Recalls "blue"	357	0.867	0.896
Acute	—	—	—
HHA	98	0.947	0.952
IRF	106	0.829	0.881
LTCH	40	0.805	0.937
SNF	113	0.852	0.834
IV.B3c.3 Recalls "bed"	357	0.858	0.914
Acute	—	—	—
HHA	98	0.888	0.929
IRF	106	0.871	0.910
LTCH	40	0.908	0.965
SNF	113	0.804	0.882
Observational Assessment of Cognitive Status			
IV.C1a Current season	19	No discordant pairs	No discordant pairs
Acute	†	†	†
HHA	†	†	†
IRF	†	†	†
LTCH	†	†	†
SNF	†	†	†
IV.C1b Location of own room	19	No discordant pairs	No discordant pairs
Acute	†	†	†
HHA	†	†	†
IRF	†	†	†
LTCH	†	†	†
SNF	†	†	†
IV.C1c Staff names and faces	19	No discordant pairs	No discordant pairs
Acute	†	†	†
HHA	†	†	†
IRF	†	†	†
LTCH	†	†	†
SNF	†	†	†
IV.C1d That he or she is in a hospital, nursing home, or home	19	0.642	N/A
Acute	†	†	N/A
HHA	†	†	N/A
IRF	†	†	N/A

(continued)

Table 2-4 (continued)
IRR testing: Cognitive status, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa
LTCH	†	†	N/A
SNF	†	†	N/A
IV.C1e None of the above are recalled	19	0.578	—
Acute	†	†	†
HHA	†	†	†
IRF	†	†	†
LTCH	†	†	†
SNF	†	†	†
IV.C1f Unable to assess	19	0.883	—
Acute	†	†	†
HHA	†	†	†
IRF	†	†	†
LTCH	†	†	†
SNF	†	†	†

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. BIMS, Brief Interview for Mental Status; HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

[a] Kappa unable to be interpreted. There were 100 agreements and 1 disagreement.

No discordant pairs and no variation in responses.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

Table 2-5
IRR testing: CAM, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa
CAM			
IV.D1 Inattention	130	0.691	0.703
Acute	†	†	†
HHA	38	0.587	0.614
IRF	36	0.743	0.815
LTCH	†	†	†
SNF	34	0.583	0.612
IV.D2 Disorganized thinking	130	0.696	0.732
Acute	†	†	†
HHA	38	0.793	0.939
IRF	36	0.683	0.686
LTCH	†	†	†
SNF	34	0.652	0.631
IV.D3 Altered level of consciousness/alertness	130	0.584	0.558
Acute	†	†	†
HHA	38	0.537	0.607
IRF	36	0.607	0.500
LTCH	†	†	†
SNF	34	0.404	0.365
IV.D4 Psychomotor retardation	130	0.474	0.477
Acute	†	†	†
HHA	38	0.224	0.558
IRF	36	0.384	0.471
LTCH	†	†	†
SNF	34	0.350	0.153
Behavioral Signs & Symptoms			
IV.E1 Physical symptoms directed toward others (e.g., hitting, kicking, pushing)	383	0.663	N/A
Acute	—	—	—
HHA	101	0.000	N/A
IRF	115	0.663	N/A
LTCH	47	1.000	N/A
SNF	120	0.796	N/A
IV.E2 Verbal symptoms directed toward others (e.g., threatening, screaming at others)	383	0.662	N/A
Acute	—	—	—
HHA	101	0.000	N/A
IRF	115	0.796	N/A
LTCH	47	1.000	N/A
SNF	120	0.658	N/A

(continued)

Table 2-5 (continued)
IRR testing: CAM, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa
IV.E3 Other disruptive or dangerous behaviors (e.g., hitting or scratching self)	382	0.745	N/A
Acute	—	—	—
HHA	100	0.000	N/A
IRF	115	1.000	N/A
LTCH	47	0.776	N/A
SNF	120	0.000	N/A

NOTE: CAM, Confusion Assessment Method; HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

Table 2-6
IRR testing: Mood and pain, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
Mood						
IV.F1 Mood interview attempted?	383	0.763	N/A	—	—	—
Acute	—	—	—	—	—	—
HHA	101	0.712	N/A	—	—	—
IRF	115	0.767	N/A	—	—	—
LTCH	47	1.000	N/A	—	—	—
SNF	120	0.600	N/A	—	—	—
PHQ-2						
IV.F2a Little interest or pleasure in doing things	328	0.860	0.856	317	0.866	N/A
Acute	—	—	—	—	—	—
HHA	94	0.908	0.753	93	0.757	N/A
IRF	86	0.868	0.921	84	0.901	N/A
LTCH	41	0.895	0.797	38	0.895	N/A
SNF	107	0.739	0.880	102	0.902	N/A
IV.F2b Number of days in the last 2 weeks (little interest or pleasure in doing things)	98	0.809	0.887	—	—	—
Acute	—	—	—	—	—	—
HHA	21	0.744	0.732	—	—	—
IRF	32	0.769	0.926	—	—	—

(continued)

Table 2-6 (continued)
IRR testing: Mood and pain, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
LTCH	19	1.000	1.000	—	—	—
SNF	26	0.716	0.811	—	—	—
IV.F2c Feeling down, depressed, or hopeless	328	0.844	0.841	317	0.841	N/A
Acute	—	—	—	—	—	—
HHA	94	0.813	N/A	94	0.813	N/A
IRF	86	0.888	0.909	83	0.876	N/A
LTCH	41	0.868	0.800	38	0.895	N/A
SNF	107	0.816	0.816	102	0.811	N/A
IV.F2d Number of days feeling down, depressed, or hopeless	112	0.849	0.907	—	—	—
Acute	—	—	—	—	—	—
HHA	29	0.752	0.741	—	—	—
IRF	32	0.952	0.980	—	—	—
LTCH	18	1.000	1.000	—	—	—
SNF	33	0.680	0.861	—	—	—
IV.F3 Patient reports feeling sad	328	0.742	0.842	—	—	—
Acute	—	—	—	—	—	—
HHA	94	0.706	0.837	94	0.706	0.837
IRF	86	0.839	0.873	83	0.828	0.826
LTCH	41	0.813	0.943	38	0.791	0.911
SNF	107	0.637	0.760	103	0.627	0.758
Pain Interview						
IV.G1 Interview attempted	449	0.630	N/A	—	—	—
Acute	65	0.488	N/A	—	—	—
HHA	101	0.795	N/A	—	—	—
IRF	115	0.696	N/A	—	—	—
LTCH	48	0.455	N/A	—	—	—
SNF	120	0.616	N/A	—	—	—
IV.G2 Pain presence: Pain during the last 2 days?	406	0.864	0.824	398	0.880	N/A
Acute	62	0.937	0.942	61	0.934	N/A
HHA	98	0.887	0.889	97	0.913	N/A
IRF	106	0.949	0.949	106	0.949	N/A
LTCH	42	0.904	0.811	38	0.934	N/A
SNF	98	0.686	0.569	96	0.715	N/A
IV.G3 Pain severity: Worst pain during the last 2 days on a zero to 10 scale	270	0.820	0.868	217	0.832	0.910
Acute	32	0.886	0.705	26	0.898	0.980
HHA	73	0.861	0.903	60	0.864	0.974
IRF	79	0.822	0.897	66	0.850	0.855
LTCH	27	0.914	0.994	18	0.933	0.994
SNF	59	0.672	0.759	47	0.671	0.829

(continued)

Table 2-6 (continued)
IRR testing: Mood and pain, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
IV.G4 Pain has an effect on sleep	265	0.829	0.836	263	0.825	N/A
Acute	31	0.868	—	31	0.868	N/A
HHA	71	0.733	—	71	0.733	N/A
IRF	79	0.825	0.836	78	0.818	N/A
LTCH	27	1.000	1.000	27	1.000	N/A
SNF	57	0.830	0.846	56	0.821	N/A
IV.G5 Pain has an effect on activities	266	0.804	0.789	261	0.820	N/A
Acute	32	1.000	1.000	32	1.000	N/A
HHA	71	0.756	0.763	70	0.781	N/A
IRF	79	0.840	0.853	77	0.822	N/A
LTCH	27	1.000	1.000	27	1.000	N/A
SNF	57	0.636	0.562	55	0.686	N/A
Pain Observational Assessment						
IV.G6a Non-verbal sounds	453	0.663	N/A	—	—	—
Acute	66	1.000	N/A	—	—	—
HHA	102	1.000	N/A	—	—	—
IRF	115	1.000	N/A	—	—	—
LTCH	49	1.000	N/A	—	—	—
SNF	121	0.388	N/A	—	—	—
IV.G6b Vocal complaints of pain	453	0.610	N/A	—	—	—
Acute	66	1.000	N/A	—	—	—
HHA	102	1.000	N/A	—	—	—
IRF	115	0.663	N/A	—	—	—
LTCH	49	1.000	N/A	—	—	—
SNF	121	0.483	N/A	—	—	—
IV.G6c Facial expressions	453	0.659	N/A	—	—	—
Acute	66	0.000[a]	N/A	—	—	—
HHA	102	1.000	N/A	—	—	—
IRF	115	0.796	N/A	—	—	—
LTCH	49	0.645	N/A	—	—	—
SNF	121	0.559	N/A	—	—	—
IV.G6d Protective body movements/postures	453	0.420	N/A	—	—	—
Acute	66	1.000	N/A	—	—	—
HHA	102	0.662	N/A	—	—	—
IRF	115	0.493	N/A	—	—	—
LTCH	49	0.021[b]	N/A	—	—	—
SNF	121	0.392	N/A	—	—	—

(continued)

Table 2-6 (continued)
IRR testing: Mood and pain, PAC admission and acute discharges, IRR sample, by provider type (CARE Tool Section 4)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
IV.G6e None of the above	453	0.643	N/A	—	—	—
Acute	66	0.792	N/A	—	—	—
HHA	102	0.390	N/A	—	—	—
IRF	115	0.483	N/A	—	—	—
LTCH	49	0.693	N/A	—	—	—
SNF	121	0.667	N/A	—	—	—

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; PHQ-2, Patient Health Questionnaire; SNF, skilled nursing facility.

[a] Kappa unable to be interpreted. There were 65 agreements and 1 disagreement.

[b] Kappa unable to be interpreted. There were 47 agreements and 2 disagreements.

* With unable to answer or no response excluded. With missings excluded.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

2.4 Section V. Impairments

Impairment items are important measures of patient severity and predictors of patient resource utilization. According to the disablement model developed by Nagi (1965), *impairment* is defined as any loss or abnormality of anatomic, physiologic, mental, or emotional structure or function. These may or may not result in functional performance limitations. This section of the CARE tool has seven individual subsections to measure impairments in bladder and bowel management, swallowing, hearing/vision/communication, weight-bearing restrictions, grip strength, respiratory status, and endurance. Each section has its own unique screening item for each type of impairment followed by the supplemental item or items measuring impairment level for those with an impairment (as noted in the screening item). Kappas reported for the screening items below apply to the full interrater reliability sample; kappas for the supplemental items are only for the segment of patients who were reported to have that type of impairment.

Table 2-7 shows interrater reliability testing results for impairments in bowel and bladder management, and ***Table 2-8*** shows results for swallowing. Bladder and bowel management can be predictive of resource utilization and outcomes. A patient with frequent incontinence and need for assistance in managing these impairments will require more resources. A patient's ability to swallow is also predictive of resource utilization and may affect post-acute care discharge options. Dysphagia, or difficulty with swallowing, is associated with increased morbidity and mortality. The swallowing impairment signs and symptoms items are based on input from the American Speech Language Hearing Association and asks the assessor to identify signs and symptoms of a possible swallowing disorder including complaints of difficulty or pain with swallowing, coughing or choking during meals, holding food in mouth, or loss of liquids or

solids from mouth when eating and drinking, and no food intake by mouth. Results have been reported for all responses (see the first three columns) and for responses excluding "not applicable" codes (in the second three columns).

The bowel and bladder items show substantial consistency between raters and across provider types: overall kappas range from 0.60 to 0.90, with most items over 0.70, showing at least substantial reliability. Kappas appear to be a bit higher for bladder items, though bowel management kappas may have been impacted by lower prevalence of impairments in bowel management. For item "Any impairment" (V.A1), acute care hospitals had the highest agreement, but they showed much lower agreement on other items. For items V.A4a and V.A4b ("Need assistance to manage equipment"), LTCHs and SNFs exhibited low kappas and the kappa for IRFs could not be interpreted, but this could be due to the inclusion of not applicable or unknown responses in the analysis.

Swallowing signs and symptoms had more variation in scores, with near-perfect overall agreement for "Intake not by mouth" (V.B1e) at 0.97. In contrast, "Complaints of difficulty swallowing" had the lowest overall kappa score in this group at 0.46, indicating moderate agreement. "Holding food in mouth" and "Loss of liquids" had overall scores of 0.56 and 0.57, respectively, indicating moderate reliability. "Coughing or choking" and "Other signs and symptoms" showed substantial agreement and raters were almost perfect when evaluating if a patient had no signs or symptoms (0.84). For the swallowing item indicating the presence or absence of any signs and symptoms of a swallowing disorder (V.B1g, "None"), unweighted kappas for acute hospitals, IRFs, and SNFs indicate almost perfect agreement whereas HHAs and LTCHs demonstrate substantial agreement. Weighted kappas were not applicable because there are only two response categories for the variable. For items V.B2a through V.B2c, relating to whether the patient can swallow regular food, modified food, or requires a feeding tube, the sample sizes were too small due to skip patterns to calculate provider-type specific kappas.

Hearing, Vision, and Communication Comprehension

The hearing, vision, and communication comprehension items on the CARE tool include four items building on the MDS 3.0. The goal of these items is to identify the level of impairment as mild or moderately impaired, severely impaired, or not impaired. Levels of impairment are assessed with hearing aids, glasses, or other assistive devices that the beneficiaries may use. These items indicate the presence or absence of a problem and the identification of a problem can lead to further assessment. These items are included in the tool because they are predictive of resource utilization and are important to communicate during care transitions. These items are shown in ***Table 2-9***. The overall kappa statistics for these are all substantial at 0.6 or higher; provider-type specific kappas range from 0.5 to 0.94, with most above 0.6, showing at least moderate reliability.

Weight-Bearing

The weight-bearing items shown in ***Table 2-9*** measure whether a patient has restrictions on their ability to bear weight in the left upper extremity, right upper extremity, left lower extremity, and right lower extremity. Restrictions on the ability to bear weight are important to capture because they are related to a patient's ability to use assistive devices and need for assistance in performing surface-to-surface transfers. This item is predictive of resource

utilization and may also be predictive of post-acute care discharge options because patients with restrictions on their ability to bear weight may require significant levels of assistance. These items showed substantial or greater consistency. In some settings, the sample sizes were too small to calculate a kappa statistic, but IRFs and SNFs had usable sample sizes for all questions in this section and had kappas ranging from 0.68 to 1.00, showing substantial to almost perfect reliability.

Grip Strength

The grip strength item measures a patient's ability to squeeze a caregiver's hand with each of their own hands. Response categories include normal, reduced/limited, or absent. This item is included in the tool as a measure of frailty and severity of illness. These items also showed substantial or greater consistency overall (see ***Table 2-9***). Kappas stratified by provider type were also consistently substantial or greater, with the exception of LTCHs, which had kappas of 0.49 and 0.43 on items regarding any impairment of grip strength and left hand grip strength, respectively, showing moderate agreement.

Respiratory Status

Providers were asked to report on shortness of breath or dyspnea associated with different levels of activity. Scores were assessed for those with or without supplemental oxygen (as appropriate) for patients with any respiratory impairment during the 2-day assessment period. Identifying the level of activity which causes or contributes to a patient being out of breath is predictive of patient severity of illness and potential resource utilization. If patients had no respiratory impairment, the level of activity item was skipped. If patients were not using supplemental oxygen, the item is entered as not applicable, likewise for patients on supplemental oxygen who would not be taken off oxygen for safety reasons. Overall weighted kappas ranged from 0.79 to 0.87 (substantial to near-perfect reliability) for items requesting levels of impairment with and without oxygen, indicating very high to almost perfect consistency between raters. LTCHs exhibited a slightly lower, but still moderate level of agreement on item "Without supplemental oxygen" (V.F1b). Kappas from prior analyses of a similar item on the OASIS ranged from 0.49 to 0.82 across several studies (Hittle et al., 2002; Berg, 1999; Madigan and Fortinsky, 2004), suggesting these results were equal to or better than past efforts in this area.

Endurance

The results for the two endurance items included on the CARE tool are also shown in ***Table 2-9***. The first is mobility endurance, which asks if the patient is able to walk or wheel 50 feet in the 2-day assessment period. The second item is sitting endurance, which asks if the patient is able to tolerate sitting for 15 minutes. Endurance is important to capture in the CARE tool because patients with low endurance are unlikely to be discharged to a rehabilitation setting where treatment includes a minimum of 15 hours of physical therapy per week. This item will be used to predict resource utilization and post-acute care discharge options. Overall kappas for both items showed substantial agreement (0.63–0.77). LTCHs again showed slightly lower but still moderate agreement on the mobility and sitting items, and IRFs showed fair agreement on item V.G1, any impairments of endurance.

Summary

This section described the reliability of items in the Impairments section of the CARE tool.

- Bowel and Bladder: Kappa scores for each item when evaluated overall ranged from 0.626 to 0.896, indicating substantial to almost perfect agreement. Agreement was substantial for most provider types and items, with a few exceptions. The two items identifying need for assistance in managing bowel and bladder equipment had lower provider type specific kappas. Unweighted kappas were in the moderate range for frequency of bladder incontinence measured in acute hospitals and HHAs. Frequency of bowel incontinence measured in IRFs and acute hospitals were lower when valid non-responses like "unknown" were kept in the analysis. Items evaluating prior use of bowel and bladder devices also had lower agreement in acute and IRFs.

- Swallowing: Results for swallowing were mixed. Agreement was perfect for the item indicating intake not by mouth (V.B1e) and the item indicated whether there were any signs or symptoms (V.B1g) had substantial to almost perfect kappas. However, agreement rates were lower and more variable across provider types for the other signs and symptoms items. SNFs had moderate to substantial kappas (0.528–0.714) for these items, and LTCHs had perfect agreement for two of the four items. Kappas were more variable for the other settings.

- Hearing, Vision, and Communication: Kappa scores for each item overall (averaged across all settings) ranged from 0.661 to 0.847, indicating substantial to almost perfect reliability. The vast majority of items had provider-type specific kappas above 0.6. The lowest kappa still indicates moderate reliability, at 0.470. In general, LTCHs and acute hospitals had the highest reliability.

- Weight Bearing: Kappa scores for each item overall (evaluated across all settings) ranged from 0.712 to 0.900, indicating substantial to almost perfect reliability. IRF and SNFs had substantial to perfect agreement on the items asking which extremity had restrictions, and were the only provider types with sufficient sample size to evaluate these items.

- Grip Strength: Kappa scores for each item overall (averaged across all settings) ranged from 0.727 to 0.885. Only LTCHs had any items with kappas scoring less than 0.6. Excluding LTCHs, the range was 0.625 to 1.00; including LTCHs, the range was 0.430 to 1.00.

- Respiratory: Kappa scores for each item overall (averaged across all settings) ranged from 0.696 to 0.874. The majority of provider types had moderate unweighted kappas but when the "not applicable" and "unknown" responses were excluded from the sample, provider types with sufficient sample size to calculate kappa had substantial to almost perfect agreement.

- Endurance: Kappa scores for each item overall (averaged across all settings) ranged from 0.539 to 0.665. Kappas were consistently substantial to almost perfect for acute and SNF settings, but consistently lower for IRFs and LTCHs.

Table 2-7
IRR testing: Bowel and bladder impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Kappa*
Impairments with Bladder and Bowel Management						
V.A1 Any impairment	452	0.844	N/A	—	—	—
Acute	66	0.905	N/A	—	—	—
HHA	102	0.874	N/A	—	—	—
IRF	115	0.770	N/A	—	—	—
LTCH	49	0.764	N/A	—	—	—
SNF	120	0.834	N/A	—	—	—
Bladder						
V.A2a External or indwelling device	251	0.896	N/A	—	—	—
Acute	25	0.754	N/A	—	—	—
HHA	61	0.782	N/A	—	—	—
IRF	69	0.910	N/A	—	—	—
LTCH	40	0.931	N/A	—	—	—
SNF	56	0.852	N/A	—	—	—
V.A3a Frequency of incontinence	251	0.711	0.831	153	0.668	0.792
Acute	25	0.541	0.768	†	†	†
HHA	61	0.579	0.757	59	0.550	0.715
IRF	69	0.727	0.865	35	0.681	0.808
LTCH	40	0.644	0.661	†	†	†
SNF	56	0.750	0.840	39	0.717	0.805
V.A4a Need assistance to manage equipment	251	0.702	N/A	—	—	—
Acute	25	0.503	N/A	—	—	—
HHA	61	0.769	N/A	—	—	—
IRF	69	0.022[a]	N/A	—	—	—
LTCH	40	0.362	N/A	—	—	—
SNF	56	0.247	N/A	—	—	—
V.A5a Incontinent/device prior	251	0.694	0.602	189	0.755	N/A
Acute	25	0.438	—	19	0.617	N/A
HHA	61	0.686	—	61	0.686	N/A
IRF	69	0.443	—	48	0.654	N/A
LTCH	40	0.777	—	†	†	N/A
SNF	56	0.824	—	47	0.828	N/A
Bowel						
V.A2b External or indwelling device	251	0.761	N/A	—	—	—
Acute	25	0.648	N/A	—	—	—
HHA	61	0.734	N/A	—	—	—
IRF	69	0.850	N/A	—	—	—
LTCH	40	0.714	N/A	—	—	—
SNF	56	0.791	N/A	—	—	—
V.A3b Frequency of incontinence	251	0.733	0.729	233	0.751	0.797
Acute	25	0.556	0.363	21	0.654	0.681
HHA	61	0.821	0.787	59	0.841	0.862

(continued)

Table 2-7 (continued)
IRR testing: Bowel and bladder impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
IRF	69	0.630	0.739	64	0.571	0.613
LTCH	40	0.611	0.706	34	0.631	0.859
SNF	56	0.842	0.846	55	0.836	0.824
V.A4b Need assistance to manage equipment	251	0.768	N/A	—	—	—
Acute	25	0.595	N/A	—	—	—
HHA	61	0.812	N/A	—	—	—
IRF	69	0.710	N/A	—	—	—
LTCH	40	−0.026[b]	N/A	—	—	—
SNF	56	0.567	N/A	—	—	—
V.A5b Incontinent/device prior	251	0.673	0.626	191	0.762	N/A
Acute	25	0.438	0.300	22	0.545	N/A
HHA	61	0.686	0.661	60	0.804	N/A
IRF	69	0.443	0.249	49	0.500	N/A
LTCH	40	0.634	0.648	†	†	N/A
SNF	56	0.875	0.874	48	0.909	N/A

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 455 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

[a] Kappa unable to be interpreted. There were 65 agreements and 4 disagreements.

[b] Kappa unable to be interpreted. There were 38 agreements and 2 disagreements.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

Table 2-8
IRR testing: Swallowing impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
Swallowing						
V.B1a Complaints of difficulty or pain with swallowing	452	0.462	N/A	—	—	—
Acute	65	0.000[a]	65	—	—	—
HHA	102	0.391	102	—	—	—
IRF	115	0.264	115	—	—	—
LTCH	49	0.000[b]	49	—	—	—
SNF	121	0.640	121	—	—	—
V.B1b Coughing or choking during meals or when swallowing medications	452	0.676	N/A	—	—	—
Acute	65	0.816	N/A	—	—	—
HHA	102	0.591	N/A	—	—	—
IRF	115	0.483	N/A	—	—	—
LTCH	49	0.790	N/A	—	—	—
SNF	121	0.714	N/A	—	—	—
V.B1c Patient holds food in mouth/cheeks or residual food in mouth after meals	452	0.562	N/A	—	—	—
Acute	65	0.000	N/A	—	—	—
HHA	102	0.662	N/A	—	—	—
IRF	115	0.659	N/A	—	—	—
LTCH	49	1.000	N/A	—	—	—
SNF	121	0.528	N/A	—	—	—
V.B1d Patient loses liquids and/or solids from mouth when eating or drinking	452	0.568	N/A	—	—	—
Acute	65	1.000	N/A	—	—	—
HHA	102	0.000	N/A	—	—	—
IRF	115	0.000	N/A	—	—	—
LTCH	49	1.000	N/A	—	—	—
SNF	121	0.663	N/A	—	—	—
V.B1e Food intake not by mouth	452	0.971	N/A	—	—	—
Acute	65	1.000	N/A	—	—	—
HHA	102	1.000	N/A	—	—	—
IRF	115	1.000	N/A	—	—	—
LTCH	49	0.918	N/A	—	—	—
SNF	121	1.000	N/A	—	—	—
V.B1f Other	452	0.646	N/A	—	—	—
Acute	65	0.484	N/A	—	—	—
HHA	102	0.658	N/A	—	—	—
IRF	115	0.781	N/A	—	—	—
LTCH	49	0.087	N/A	—	—	—
SNF	121	0.919	N/A	—	—	—

(continued)

Table 2-8 (continued)
IRR testing: Swallowing impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Excluded kappa*	Weighted kappa*
V.B1g None	452	0.839	N/A	—	—	—
Acute	65	0.882	N/A	—	—	—
HHA	102	0.649	N/A	—	—	—
IRF	115	0.922	N/A	—	—	—
LTCH	49	0.671	N/A	—	—	—
SNF	121	0.880	N/A	—	—	—
V.B2a Patient can swallow regular food	15	1.000	—	—	—	—
Acute	†	†	—	—	—	—
HHA	†	†	—	—	—	—
IRF	†	†	—	—	—	—
LTCH	†	†	—	—	—	—
SNF	†	†	—	—	—	—
V.B2b Patient can swallow modified food	15	1.000	—	—	—	—
Acute	†	†	—	—	—	—
HHA	†	†	—	—	—	—
IRF	†	†	—	—	—	—
LTCH	†	†	—	—	—	—
SNF	†	†	—	—	—	—
V.B2c Patient requires tube or parenteral feeding wholly or partially	15	1.000	—	—	—	—
Acute	†	†	—	—	—	—
HHA	†	†	—	—	—	—
IRF	†	†	—	—	—	—
LTCH	†	†	—	—	—	—
SNF	†	†	—	—	—	—

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

[a] Kappa unable to be interpreted. There were 64 agreements and 1 disagreement.

[b] Kappa unable to be interpreted. There were 48 agreements and 1 disagreement.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

Table 2-9
IRR testing: Other impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
Hearing, Vision, or Communication						
V.C1 Patient has hearing, vision, or communication impairment(s)	453	0.769	N/A	—	—	—
Acute	66	0.848	N/A	—	—	—
HHA	102	0.854	N/A	—	—	—
IRF	115	0.609	N/A	—	—	—
LTCH	49	0.847	N/A	—	—	—
SNF	121	0.742	N/A	—	—	—
V.C1a Patient understands verbal content	219	0.693	0.728	206	0.677	0.777
Acute	28	0.740	0.535	24	0.694	0.858
HHA	60	0.470	0.498	58	0.425	0.398
IRF	56	0.642	0.779	54	0.648	0.772
LTCH	34	0.796	0.758	30	0.798	0.895
SNF	41	0.853	0.925	40	0.845	0.915
V.C1b Patient can express needs and wants	219	0.661	0.713	208	0.656	0.789
Acute	28	0.686	0.696	26	0.694	0.902
HHA	60	0.610	0.668	59	0.594	0.643
IRF	56	0.539	0.635	54	0.534	0.716
LTCH	34	0.809	0.801	30	0.817	0.910
SNF	41	0.627	0.628	39	0.633	0.642
V.C1c Ability to see in adequate light	219	0.743	0.780	201	0.744	0.748
Acute	28	0.942	0.965	25	0.925	0.939
HHA	60	0.739	0.691	59	0.726	0.667
IRF	56	0.554	0.542	52	0.663	0.696
LTCH	34	0.863	0.930	25	0.916	0.937
SNF	41	0.651	0.744	40	0.628	0.715
V.C1d Ability to hear	219	0.780	0.838	206	0.763	0.800
Acute	28	1.000	1.000	25	1.000	1.000
HHA	60	0.648	0.725	59	0.628	0.671
IRF	56	0.678	0.656	55	0.701	0.753
LTCH	34	0.884	0.922	27	0.867	—
SNF	41	0.705	0.796	40	0.687	0.762
Weight Bearing						
V.D1 Any impairments with weight bearing	450	0.760	N/A	—	—	—
Acute	66	1.000	N/A	—	—	—
HHA	102	0.587	N/A	—	—	—
IRF	115	0.716	N/A	—	—	—

(continued)

Table 2-9 (continued)
IRR testing: Other impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
LTCH	46	0.849	N/A	—	—	—
SNF	121	0.826	N/A	—	—	—
V.D1a Upper left extremity	60	0.763	N/A	—	—	—
Acute	†	†	N/A	—	—	—
HHA	†	†	N/A	—	—	—
IRF	19	0.759	N/A	—	—	—
LTCH	†	†	N/A	—	—	—
SNF	22	0.776	N/A	—	—	—
V.D1b Upper right extremity	60	0.712	N/A	—	—	—
Acute	†	†	N/A	—	—	—
HHA	†	†	N/A	—	—	—
IRF	19	0.683	N/A	—	—	—
LTCH	†	†	N/A	—	—	—
SNF	22	1.000	N/A	—	—	—
V.D1c Lower left extremity	60	0.900	N/A	—	—	—
Acute	†	†	N/A	—	—	—
HHA	†	†	N/A	—	—	—
IRF	19	0.895	N/A	—	—	—
LTCH	†	†	N/A	—	—	—
SNF	22	0.909	N/A	—	—	—
V.D1d Lower right extremity	60	0.798	N/A	—	—	—
Acute	†	†	N/A	—	—	—
HHA	†	†	N/A	—	—	—
IRF	19	0.671	N/A	—	—	—
LTCH	†	†	N/A	—	—	—
SNF	22	0.820	N/A	—	—	—
Grip Strength						
V.E1 Any impairments of grip strength	449	0.766	N/A	—	—	—
Acute	66	0.891	N/A	—	—	—
HHA	102	0.778	N/A	—	—	—
IRF	115	0.625	N/A	—	—	—
LTCH	45	0.492	N/A	—	—	—
SNF	121	0.821	N/A	—	—	—
V.E1a Left hand	103	0.752	0.813	—	—	—
Acute	†	†	†	—	—	—
HHA	†	†	†	—	—	—
IRF	32	0.871	0.894	—	—	—
LTCH	31	0.430	0.530	—	—	—
SNF	17	1.000	1.000	—	—	—

(continued)

Table 2-9 (continued)
IRR testing: Other impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
V.E1b Right hand	103	0.853	0.885	—	—	—
Acute	†	†	†	—	—	—
HHA	†	†	†	—	—	—
IRF	32	0.739	0.828	—	—	—
LTCH	31	0.874	0.884	—	—	—
SNF	17	1.000	1.000	—	—	—
Respiratory Status						
V.F1 Any respiratory impairments	453	0.815	N/A	—	—	—
Acute	66	0.593	N/A	—	—	—
HHA	102	0.824	N/A	—	—	—
IRF	116	0.691	N/A	—	—	—
LTCH	48	0.846	N/A	—	—	—
SNF	121	0.920	N/A	—	—	—
V.F1a With supplemental oxygen	145	0.727	0.859	64	0.617	0.791
Acute	†	†	†	†	†	†
HHA	46	0.455	0.951	†	†	†
IRF	27	0.545	0.469	†	†	†
LTCH	33	0.794	0.769	†	†	†
SNF	32	0.807	0.974	24	0.826	0.902
V.F1b Without supplemental oxygen	145	0.696	0.874	79	0.620	0.815
Acute	†	†	†	†	†	†
HHA	47	0.578	0.908	40	0.509	0.708
IRF	27	0.567	0.764	15	0.605	0.801
LTCH	33	0.516	0.579	†	†	†
SNF	32	0.766	0.971	18	0.773	0.933
Endurance						
V.G1 Any impairments of endurance	448	0.605	N/A	—	—	—
Acute	64	0.658	N/A	—	—	—
HHA	102	0.578	N/A	—	—	—
IRF	115	0.320	N/A	—	—	—
LTCH	46	0.657	N/A	—	—	—
SNF	121	0.654	N/A	—	—	—
V.G1a Mobility	327	0.694	0.665	276	0.713	0.768
Acute	25	0.765	0.686	22	0.776	0.853
HHA	79	0.601	0.650	71	0.586	0.681
IRF	85	0.590	0.749	84	0.579	0.730
LTCH	44	0.316	0.092	†	†	†
SNF	94	0.902	0.893	86	0.923	0.918

(continued)

Table 2-9 (continued)
IRR testing: Other impairments at PAC admission and acute discharge, IRR sample, by provider type (CARE Tool Section 5)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
V.G1b Sitting	327	0.635	0.539	297	0.628	0.699
Acute	25	0.732	0.612	22	0.725	0.784
HHA	79	0.664	0.584	76	0.682	0.738
IRF	85	0.386	0.492	84	0.360	0.412
LTCH	44	0.427	0.374	25	0.443	0.728
SNF	94	0.758	0.794	90	0.752	0.746

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

2.5 Section VI. Functional Status

2.5.1 Core Function Items

The CARE tool includes a core set of six self-care items and five functional mobility items that are asked of all patients. Items represent a range of difficulty. Many of these are modified from existing items on the OASIS, MDS 3.0, and IRF-PAI. The primary purpose of each of the function items is to understand the potential resource utilization and post-acute care discharge decisions as measured through the independence or need for assistance scale.

The core items are rated using a six-level rating scale measuring the patient's independence or need for assistance. Rating scale levels include total dependence, substantial/maximal assistance, partial/moderate assistance, supervision or touching assistance, setup or cleanup assistance, or total independence. Respondents can also indicate that the item was not attempted due to medical or safety concerns, attempted but not completed, not applicable to the patient, or the patient refused. Because these "not attempted" responses are not ordinal to each other nor were clinicians trained to differentiate finely between these responses, we are reporting a set of kappas where these responses have been set to missing. An additional analysis, not shown, was conducted where kappas were calculated using recoded function items that grouped "not attempted" responses together. Results uniformly showed only slight increases in the unweighted kappas, largely in the third decimal, and slight decreases in the weighted kappas from what is reported below.

This core set of items evaluates all patients, regardless of functional level, on basic self-care activities such as eating, tube feeding, oral hygiene, toilet hygiene, and upper- and lower-body dressing. The core mobility items include patient ability to move from lying to sitting on

the side of the bed, to move from sitting position to standing, to transfer to and from a chair (or wheelchair), and to get on and off a toilet or commode. Results for these core items are reported below in ***Table 2-10*** and are split into two conceptual groupings corresponding to self-care and mobility items.

The core mobility section of the CARE tool includes items characterizing patient's level of independence in locomotion or ambulation structured with a screening question that asks the patient's mode of mobility, or whether the patient primarily uses a wheelchair for mobility. The subsequent questions request information on the patient's level of independence in mobility at the longest distance they are able to ambulate (150, 100, or 50 feet or in room), separating responses for patients who walk from those who primarily wheel. Effective sample sizes for these items are smaller because each patient has a response for a single one of these eight modes of mobility items.

Overall kappa statistics for all core items, self-care and mobility, indicate substantial agreement among raters with the exception of three items: "Tube feeding," "Oral hygiene," and "Toilet transfer." (Note that the "Wheel 150 feet" item (VI.B5b1), the "Wheel 100 feet" item (VI.B5b2), and the "Wheel 50 feet" item (VI.B5b3) were excluded due to low sample size.) The weighted kappa values for the self-care items excluding "Tube feeding" and "Oral hygiene" range between 0.798 for eating to 0.869 for upper-body dressing.

Provider-specific analyses of the self-care and mobility items in ***Table 2-10*** below show similar levels of agreement as the overall estimates. IRFs and LTCHs appear to have slightly lower rates of agreement across items than other settings. For the "Eating" item, acute hospitals showed substantial to almost perfect agreement (0.95). Unweighted kappas for HHAs, IRFs, LTCHs, and SNFs each indicate a moderate level of agreement, with the weighted kappa for both HHAs and SNFs showing substantial agreement between raters. The unweighted kappas for HHAs, IRFs, LTCHs, and SNFs, when the "not-attempted" responses were excluded, showed a moderate level of agreement, and in each case the weighted kappa was markedly higher. SNFs exhibited almost perfect agreement and substantial agreement was observed for HHAs, IRFs, and LTCHs.

For the "Oral hygiene" item (VI.A3), unweighted kappa scores including "not attempted" responses indicate substantial agreement in both acute care hospitals and HHAs, with the weighted kappa for acute care providers indicating almost perfect agreement (0.94). The unweighted kappa for SNFs indicates moderate agreement, whereas the weighted kappa indicates almost perfect agreement among raters. Unweighted kappas for both IRFs and LTCHs indicate only fair agreement, whereas the weighted kappa indicates moderate agreement in IRFs and substantial agreement for LTCHs. In the analyses excluding "not attempted" responses, unweighted kappas for all provider types remain in the same range. The weighted kappas for acute hospitals and SNFs indicate almost perfect agreement, substantial agreement for HHAs and IRFs, and moderate agreement for LTCHs. For the "Toilet hygiene" item (VI.A4), unweighted kappa scores including "not attempted" responses indicate substantial agreement for acute care hospitals and SNFs, moderate agreement for HHAs and IRFs, and fair agreement in LTCHs. Weighted kappas are higher across the board, with almost perfect agreement for acute hospitals; substantial agreement for HHAs, LTCHs, and SNFs; and moderate agreement for IRFs. When "not attempted" responses are excluded, unweighted kappas for all provider types remain in the

same range. The weighted kappa for acute hospitals indicates almost perfect agreement; substantial agreement for HHAs, LTCHs, and SNFs; and almost perfect agreement for LTCHs.

For the "Lower-body dressing" item (VI.A6), unweighted kappa scores with "not attempted" included indicate substantial interrater agreement for acute care hospitals and moderate agreement for HHAs, IRFs, LTCHs, and SNFs. The weighted kappas indicate greater agreement for all provider types with almost perfect agreement for acute hospitals and IRFs, and substantial agreement for HHAs, LTCHs, and SNFs. When "not attempted" responses are excluded, unweighted kappas indicate substantial interrater agreement in acute care hospitals; moderate agreement for HHAs, IRFs, and SNFs; and fair agreement in LTCHs (n = 20). The weighted kappas again indicate greater agreement for all provider types with acute care hospitals and IRFs demonstrating almost perfect agreement, and HHAs, LTCHs, and SNFs demonstrating substantial agreement.

For the core mobility item, "Lying to sitting on side of bed" (VI.B1), the unweighted kappas with "not attempted" responses included indicate almost perfect agreement in SNFs, substantial agreement in HHAs, and moderate agreement in acute care hospitals, IRFs, and LTCHs. The weighted kappas for LTCHs and SNFs indicate almost perfect agreement, and for acute hospitals, HHAs, and IRFs they indicate substantial agreement. In the analyses with "not attempted" excluded, unweighted kappas indicate almost perfect agreement for SNFs; moderate agreement for acute care hospitals, HHAs, and IRFs; and fair agreement for LTCHs (n = 27).

Weighted kappa scores demonstrate almost perfect agreement in acute care hospitals and SNFs, and substantial agreement in HHAs, IRFs, and LTCHs. For the core mobility item, "Sit to stand" (VI.B2), unweighted kappas with "not attempted" responses included show almost perfect agreement in SNFs; substantial agreement in acute care hospitals, HHAs, and IRFs; and moderate agreement in LTCHs. The weighted kappas for this variable indicate almost perfect agreement among clinicians in both IRFs and SNFs, substantial agreement in acute care hospitals and HHAs, and moderate agreement in LTCHs. When "not attempted" responses excluded, the unweighted kappas again indicate almost perfect agreement in SNFs and substantial agreement in acute care hospitals, HHAs, and IRFs. Weighted kappas for acute care hospitals, HHAs, IRFs, and SNFs each indicate almost perfect agreement among clinicians in these care settings. For the "Chair/Bed-to-chair transfer" (VI.B3), unweighted kappas with "not attempted" responses included show almost perfect agreement in SNFs, substantial agreement in HHAs, and moderate agreement in acute care facilities, IRFs, and LTCHs. The weighted kappas for acute care facilities indicate almost perfect agreement; for SNFs, IRFs, and HHAs substantial agreement; and for raters in LTCHs moderate agreement. With "not attempted" responses excluded, the unweighted kappas indicate almost perfect agreement in SNFs, substantial agreement in acute care hospitals and HHAs, and moderate agreement in IRFs. The weighted kappa scores for this variable show almost perfect agreement in acute care hospitals, HHAs, and SNFs, and substantial agreement in IRFs.

At the provider level, weighted kappa values were generally in line with the reliability estimates from prior studies of comparable items on the Medicare mandated assessments (FIM®, OASIS and MDS), with a few exceptions. Results for CARE were comparable to results for FIM® and OASIS items. Compared to prior MDS studies, CARE function items had largely similar, though slightly lower kappas, except for "Chair/bed-to-chair transfer," which had

slightly better results than the MDS 3.0, and "Tube feeding," "Toilet Hygiene," and "Lower body dressing," which had lower kappas. Differences in kappas may be explained by different sample populations, data collection approaches (e.g., use of gold-standard nurses), and sample sizes. Additionally, because clinicians were required to fill out both the CARE and the currently mandated assessment, there may have been an unexpected influence on responses on the CARE tool.

Summary

This section discusses the kappas for the core function items in the CARE tool.

- Reliability was very good for these items, as all overall kappa statistics indicate substantial or higher agreement among raters. Items VI.B5b1 through VI.B5b3 (the distance wheeled items) were excluded due to low sample size.

- These kappas were similar to those found for the FIM® and OASIS items that capture similar concepts, but lower than the corresponding items on MDS 2.0 and 3.0.

- Kappas stratified by provider type show similar levels of agreement to the overall kappas. IRFs and LTCHs appear to have slightly lower rates of agreement across items than other settings.

Table 2-10
IRR testing: Core self-care and mobility at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
Core Self-Care						
VI.A1 Eating	449	0.620	0.692	401	0.617	0.798
Acute	66	0.779	0.950	64	0.763	0.943
HHA	102	0.590	0.610	102	0.590	0.610
IRF	114	0.459	0.563	104	0.469	0.726
LTCH	46	0.446	0.581	16	0.422	0.727
SNF	121	0.592	0.718	115	0.574	0.856
VI.A2 Tube feeding	450	0.594	0.890	18	0.217	0.781
Acute	66	0.662	0.887	†	†	†
HHA	102	0.124	0.487	†	†	†
IRF	115	0.735	0.885	†	†	†
LTCH	46	0.546	0.841	†	†	†
SNF	121	0.679	0.867	†	†	†
VI.A3 Oral hygiene	450	0.586	0.766	414	0.598	0.842
Acute	66	0.727	0.942	65	0.744	0.957
HHA	102	0.611	0.721	101	0.625	0.722
IRF	115	0.405	0.585	103	0.405	0.799
LTCH	46	0.331	0.705	25	0.254	0.555
SNF	121	0.587	0.875	120	0.581	0.871

(continued)

Table 2-10 (continued)
IRR testing: Core self-care and mobility at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
VI.A4 Toilet hygiene	450	0.619	0.777	416	0.636	0.845
Acute	66	0.672	0.906	66	0.672	0.906
HHA	102	0.608	0.758	102	0.608	0.758
IRF	115	0.531	0.576	105	0.556	0.738
LTCH	46	0.339	0.753	22	0.344	0.813
SNF	121	0.645	0.791	121	0.645	0.791
VI.A5 Upper-body dressing	450	0.629	0.826	420	0.634	0.869
Acute	66	0.650	0.816	61	0.698	0.929
HHA	102	0.621	0.832	101	0.629	0.846
IRF	115	0.537	0.836	114	0.531	0.825
LTCH	46	0.459	0.740	23	0.400	0.783
SNF	121	0.657	0.839	121	0.657	0.839
VI.A6 Lower-body dressing	450	0.617	0.804	413	0.625	0.855
Acute	66	0.681	0.844	60	0.724	0.925
HHA	102	0.584	0.794	101	0.591	0.806
IRF	115	0.595	0.885	112	0.590	0.861
LTCH	46	0.447	0.696	20	0.396	0.754
SNF	121	0.589	0.644	120	0.596	0.702
Core Mobility						
VI.B1 Lying to sitting on side of bed	449	0.701	0.813	412	0.693	0.855
Acute	65	0.561	0.723	61	0.580	0.861
HHA	102	0.633	0.777	97	0.600	0.734
IRF	115	0.579	0.637	109	0.595	0.796
LTCH	46	0.602	0.863	27	0.360	0.728
SNF	121	0.844	0.811	118	0.849	0.878
VI.B2 Sit to stand	449	0.752	0.814	387	0.762	0.901
Acute	65	0.622	0.724	60	0.638	0.869
HHA	102	0.620	0.727	98	0.621	0.813
IRF	115	0.717	0.843	103	0.730	0.895
LTCH	46	0.551	0.597	†	†	†
SNF	121	0.879	0.831	114	0.916	0.924
VI.B3 Chair/bed-to-chair transfer	448	0.645	0.780	392	0.752	0.901
Acute	65	0.598	0.879	62	0.610	0.861
HHA	102	0.663	0.665	95	0.744	0.855
IRF	115	0.588	0.734	110	0.583	0.788
LTCH	46	0.556	0.520	†	†	†
SNF	120	0.899	0.789	115	0.916	0.934
VI.B4 Toilet transfer	448	0.559	0.757	361	0.688	0.878
Acute	64	0.615	0.890	60	0.617	0.857
HHA	102	0.682	0.802	99	0.715	0.828

(continued)

Table 2-10 (continued)
IRR testing: Core self-care and mobility at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
IRF	115	0.397	0.742	83	0.363	0.698
LTCH	46	0.398	0.128	†	†	†
SNF	121	0.802	0.739	112	0.839	0.917
VI.B5 Wheelchair use	449	0.866	N/A	449	0.866	N/A
Acute	65	0.815	N/A	65	0.815	N/A
HHA	102	0.879	N/A	102	0.879	N/A
IRF	115	0.743	N/A	115	0.743	N/A
LTCH	46	0.867	N/A	46	0.867	N/A
SNF	121	0.934	N/A	121	0.934	N/A
VI.B5a1 Walk 150 feet	70	0.787	0.666	68	0.774	0.558
Acute	32	0.704	0.605	32	0.704	0.605
HHA	†	†	†	†	†	†
IRF	†	†	†	†	†	†
LTCH	†	†	†	†	†	†
SNF	†	†	†	†	†	†
VI.B5a2 Walk 100 feet	29	0.925	0.971	29	0.925	0.971
Acute	†	†	†	†	†	†
HHA	†	†	†	†	†	†
IRF	†	†	†	†	†	†
LTCH	†	†	†	†	†	†
SNF	10	†	†	†	†	†
VI.B5a3 Walk 50 feet	49	0.773	0.929	49	0.773	0.929
Acute	†	†	†	†	†	†
HHA	21	0.802	0.955	21	0.802	0.955
IRF	†	†	†	†	†	†
LTCH	†	†	†	†	†	†
SNF	†	†	†	†	†	†
VI.B5a4 Walk once standing	80	0.707	0.858	52	0.667	0.836
Acute	†	†	†	†	†	†
HHA	18	0.843	0.355	17	0.913	0.963
IRF	21	0.198	0.506	18	0.250	0.286
LTCH	21	0.378	0.624	†	†	†
SNF	†	†	†	†	†	†
VI.B5b1 Wheel 150 feet	†	†	†	†	†	†
Acute	†	†	†	†	†	†
HHA	†	†	†	†	†	†
IRF	†	†	†	†	†	†
LTCH	†	†	†	†	†	†
SNF	†	†	†	†	†	†

(continued)

Table 2-10 (continued)
IRR testing: Core self-care and mobility at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
VI.B5b2 Wheel 100 feet	†	†	†	†	†	†
Acute	†	†	†	†	†	†
HHA	†	†	†	†	†	†
IRF	†	†	†	†	†	†
LTCH	†	†	†	†	†	†
SNF	†	†	†	†	†	†
VI.B5b3 Wheel 50 feet	†	†	†	†	†	†
Acute	†	†	†	†	†	†
HHA	†	†	†	†	†	†
IRF	†	†	†	†	†	†
LTCH	†	†	†	†	†	†
SNF	†	†	†	†	†	†
VI.B5b4 Wheel in room	85	0.714	0.767	46	0.751	0.924
Acute	†	†	†	†	—	—
HHA	†	†	†	†	†	†
IRF	15	0.524	0.966	†	†	†
LTCH	15	0.196	0.466	†	†	†
SNF	43	0.839	0.850	28	0.955	0.993

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

*With not attempted, N/A, or refused excluded.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

2.5.2 Supplemental Function Items

Table 2-11 shows patients' level of independence in supplemental self-care items such as the ability to wash, rinse, and dry the upper body and to bathe self in the shower or tub. Overall kappas show substantial consistency when the "not attempted" responses were included and almost perfect agreement with the "not attempted" responses excluded. ***Table 2-11*** also shows supplemental mobility items such as rolling from lying on the back to left and right side, to move from sitting on side of the bed to lying flat on the bed, to bend/stoop from a standing position to pick up a small object from the floor, and the ability to put on and take off socks and shoes or other footwear. For patients whose mode of ambulation is walking, this table also shows the ability to step over a curb or up and down one step, to walk 50 feet and make two turns, to go up

and down 12 interior steps with a rail, to go up and down four exterior steps with a rail, to walk 10 feet on uneven or sloping surfaces, and to transfer in and out of a car. For patients whose mode of ambulation is wheeling, this table shows patient ability to wheel on a short ramp and on a long ramp. Supplemental mobility items showed more variability in kappa scores. Agreement is nearly perfect when excluding "not attempted" responses, but the sample is small for four steps exterior (VI.C7d), walk 10 feet on uneven surface (VI.C7e), and wheel short and long ramps (VI.C7g, VI.C7h).

Overall kappas shown in ***Table 2-11*** for instrumental activities of daily living (IADLs) generally had substantial consistency or better except for light shopping and using public transportation. Using public transportation had lower kappas when the not-assessed responses were included, but had substantial agreement when those responses were included. The opposite was true for laundry.

Provider-specific analyses of supplemental and IADL items in ***Table 2-11*** below show similar agreement to the overall estimates. Because "not attempted" responses were much more common for these items, particularly the more difficult to perform IADLs and supplemental mobility items like climb 12 stairs, there are large differences between the kappas calculated with the "not attempted" responses included. Less emphasis was placed on the choice of "not attempted" code in CARE tool trainings, likely resulting in larger variation in this type of response between clinician pairs.

For "Wash upper body" (VI.C1), unweighted kappas with "not attempted" responses excluded included indicate substantial interrater agreement in outpatient settings (HHAs and SNFs), moderate agreement in acute care hospitals and IRFs, and slight agreement in LTCHs. The weighted kappas show higher scores across the board, indicating almost perfect agreement for HHAs and SNFs; LTCHs had substantial agreement, and IRFs moderate. When "not attempted" responses are excluded, unweighted kappas indicate almost perfect agreement in SNFs, substantial agreement in acute hospitals and HHAs, and moderate agreement in IRFs. Weighted kappas present higher scores for all provider types. Acute care hospitals, HHAs, and SNFs demonstrate almost perfect agreement whereas IRFs indicate substantial agreement.

For "Roll left and right" (VI.C3), unweighted kappas with "not attempted" responses included indicate almost perfect agreement in SNFs, moderate agreement in acute hospitals and HHAs, and fair agreement in IRFs and LTCHs. Weighted kappas indicate almost perfect agreement in acute care hospitals, moderate agreement in IRFs, and fair agreement in LTCHs. Unweighted kappas with "not attempted" responses excluded indicate agreement with the following ratings: almost perfect in SNFs, substantial for acute care providers, moderate in HHAs and IRFs, and fair agreement in LTCHs. Weighted kappas indicate almost perfect agreement in acute hospitals and SNFs, substantial agreement in HHAs and IRFs, and moderate agreement in LTCHs. For "Put on/ take off footwear" (VI.C6), unweighted kappas with "not attempted" responses included indicate substantial agreement in outpatient settings, moderate agreement in acute care hospitals and IRFs, and fair agreement in LTCHs. Weighted kappas indicate almost perfect interrater agreement in HHAs; substantial agreement in acute care hospitals, LTCHs, and SNFs; and moderate agreement in IRFs. Unweighted kappas with "not attempted" responses included for acute care hospitals and SNFs indicate almost perfect agreement; HHAs had substantial agreement and IRFs moderate agreement. The weighted

kappas calculated excluding "not attempted" responses indicate almost perfect interrater agreement in acute care hospitals, HHAs, IRFs, and SNFs.

For "12 steps interior" (VI.C7c), unweighted kappas with "not attempted" responses included indicate almost perfect agreement in SNFs, moderate agreement in HHAs, and only slight agreement in acute hospitals, IRFs, and LTCHs. Weighted kappas reveal almost perfect agreement in SNFs, substantial agreement in HHAs, moderate agreement in IRFs, fair agreement in LTCHs, and slight agreement in acute hospitals. Provider-specific analyses with "not attempted" responses excluded are not reported due to small sample sizes.

For "Oral drug management" (VI.C10), unweighted kappas with "not attempted" responses included indicate substantial agreement in outpatient settings, moderate agreement in IRFs and LTCHs, and poor agreement in acute care hospitals. Weighted kappas for this variable indicate almost perfect agreement in HHAs and LTCHs, substantial agreement in SNFs, fair agreement in IRFs, and poor agreement in acute care hospitals. When "not attempted" responses are excluded, unweighted kappas indicate substantial agreement among raters in HHAs and IRFs, and fair agreement in SNFs. Weighted kappas indicate almost perfect agreement in HHAs and IRFs, and substantial agreement in SNFs.

Summary

This section discusses the reliability of the supplemental function items in the CARE tool.

- Overall kappas generally demonstrate substantial agreement when "not attempted" responses were included and almost perfect agreement when "not attempted" responses were excluded.

- Additionally, the CARE tool kappas were higher than those found in previous testing of equivalent items on the OASIS assessment.

- The provider-specific analyses showed similar agreement to the overall estimates in many cases. The kappas for IRFs are consistently lower than other provider types on many items but generally increase when "not attempted" responses are excluded.

Table 2-11
IRR testing: Function—supplemental self-care, mobility, and IADLs at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
Supplemental Self-Care						
VI.C1 Wash upper body	404	0.611	0.695	353	0.638	0.861
Acute	36	0.508	0.554	27	0.626	0.947
HHA	94	0.646	0.815	93	0.657	0.814
IRF	115	0.425	0.527	103	0.430	0.745
LTCH	38	0.179	0.708	†	†	†
SNF	121	0.792	0.820	116	0.813	0.944
VI.C2 Shower/bathe self	404	0.611	0.675	254	0.625	0.867
Acute	36	0.492	0.167	20	0.755	0.958
HHA	94	0.556	0.782	90	0.558	0.805
IRF	115	0.452	0.504	62	0.458	0.807
LTCH	38	0.427	0.815	†	†	†
SNF	121	0.813	0.846	69	0.793	0.943
VI.C3 Roll left/right	402	0.614	0.579	362	0.657	0.843
Acute	36	0.591	0.843	35	0.611	0.873
HHA	92	0.540	0.721	90	0.528	0.703
IRF	115	0.400	0.446	93	0.444	0.795
LTCH	38	0.321	0.320	26	0.332	0.517
SNF	121	0.811	0.784	118	0.826	0.857
VI.C4 Sit to lying	403	0.655	0.630	350	0.711	0.857
Acute	36	0.668	0.737	33	0.702	0.876
HHA	93	0.605	0.578	87	0.651	0.788
IRF	115	0.430	0.401	91	0.520	0.763
LTCH	38	0.496	0.778	21	0.309	0.331
SNF	121	0.826	0.675	118	0.853	0.849
VI.C5 Pick up object	402	0.391	0.649	166	0.747	0.804
Acute	36	0.518	0.773	†	†	†
HHA	93	0.474	0.347	76	0.584	0.739
IRF	115	0.327	0.338	16	0.830	0.975
LTCH	38	0.342	0.708	†	†	†
SNF	120	0.769	0.830	56	0.864	0.861
VI.C6 Put on/take off footwear	400	0.652	0.738	322	0.724	0.898
Acute	35	0.544	0.778	15	0.917	0.989
HHA	92	0.695	0.830	88	0.719	0.903
IRF	115	0.474	0.580	103	0.508	0.837
LTCH	38	0.357	0.648	†	†	†
SNF	120	0.805	0.788	104	0.862	0.872

(continued)

Table 2-11 (continued)
IRR testing: Function—supplemental self-care, mobility, and IADLs at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
VI.C7 Primary wheelchair use	404	0.833	N/A	404	0.833	N/A
Acute	36	0.839	N/A	36	0.839	N/A
HHA	94	0.832	N/A	94	0.832	N/A
IRF	115	0.643	N/A	115	0.643	N/A
LTCH	38	0.892	N/A	38	0.892	N/A
SNF	121	0.934	N/A	121	0.934	N/A
VI.C7a One-step curb	242	0.510	0.702	59	0.648	0.806
Acute	26	0.209	−0.061[a]	†	†	†
HHA	77	0.489	0.564	39	0.625	0.827
IRF	64	0.259	0.785	†	†	†
LTCH	21	0.099	0.233	†	†	†
SNF	54	0.855	0.889	†	†	†
VI.C7b Walk 50 feet with two turns	242	0.513	0.535	112	0.748	0.887
Acute	26	0.098	−0.056[a]	†	†	†
HHA	77	0.515	0.474	53	0.666	0.853
IRF	64	0.397	0.537	26	0.634	0.820
LTCH	21	0.071	0.227	†	†	†
SNF	54	0.732	0.805	26	0.907	0.959
VI.C7c Twelve steps/interior	242	0.499	0.667	15	0.696	0.949
Acute	26	0.167	0.072	†	†	†
HHA	77	0.543	0.717	†	†	†
IRF	64	0.184	0.502	†	†	†
LTCH	21	0.050	0.396	†	†	†
SNF	54	0.869	0.913	†	†	†
VI.C7d Four steps/exterior	241	0.459	0.631	26	0.723	0.946
Acute	25	0.136	0.056	†	†	†
HHA	77	0.432	0.568	13	†	†
IRF	64	0.192	0.359	†	†	†
LTCH	21	0.082	−0.055[b]	†	†	†
SNF	54	0.882	0.923	†	†	†
VI.C7e Walk 10 feet on uneven surface	242	0.485	0.581	27	0.782	0.947
Acute	26	0.155	−0.049[a]	†	†	†
HHA	77	0.347	0.486	17	0.764	0.957
IRF	64	0.369	0.676	6	†	†
LTCH	21	0.012	0.134	†	†	†
SNF	54	0.842	0.911	†	†	†

(continued)

Table 2-11 (continued)
IRR testing: Function—supplemental self-care, mobility, and IADLs at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
VI.C7f Car transfer	400	0.523	0.652	80	0.773	0.926
Acute	35	0.157	0.006	†	†	†
HHA	92	0.435	0.570	45	0.658	0.907
IRF	115	0.303	0.763	†	†	†
LTCH	38	0.652	0.533	†	†	†
SNF	120	0.722	0.658	28	0.950	0.985
VI.C7g Wheel short ramp	128	0.616	0.362	†	†	†
Acute	†	†	†	†	†	†
HHA	†	†	†	†	†	†
IRF	32	0.391	0.475	†	†	†
LTCH	15	0.528	0.389	†	†	†
SNF	63	0.774	0.546	†	†	†
VI.C7h Wheel long ramp	128	0.605	0.369	†	†	†
Acute	†	†	†	†	—	—
HHA	†	†	†	†	†	†
IRF	32	0.391	0.475	†	†	†
LTCH	15	0.528	0.389	†	—	—
SNF	63	0.749	0.546	†	†	†
VI.C8 Telephone answering	402	0.611	0.622	273	0.671	0.806
Acute	34	0.332	0.468	†	†	†
HHA	94	0.607	0.654	89	0.658	0.830
IRF	115	0.377	0.396	64	0.423	0.778
LTCH	38	0.558	0.810	†	†	†
SNF	121	0.802	0.632	100	0.847	0.914
VI.C9 Telephone placing call	402	0.623	0.609	269	0.718	0.812
Acute	34	0.363	0.561	†	†	†
HHA	94	0.701	0.656	88	0.772	0.888
IRF	115	0.368	0.320	62	0.425	0.771
LTCH	38	0.527	0.816	†	†	†
SNF	121	0.815	0.645	99	0.878	0.847
VI.C10 Oral drug management	403	0.595	0.732	153	0.592	0.813
Acute	35	0.005	—	†	†	†
HHA	94	0.679	0.869	92	0.682	0.866
IRF	115	0.497	0.229	15	0.706	0.868
LTCH	38	0.489	0.883	†	†	†
SNF	121	0.622	0.756	33	0.405	0.627
VI.C11 Inhalant drug management	403	0.479	0.654	52	0.443	0.727
Acute	36	−0.006[a]	0.095	†	†	†
HHA	93	0.509	0.726	17	0.480	0.804

(continued)

Table 2-11 (continued)
IRR testing: Function—supplemental self-care, mobility, and IADLs at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
IRF	115	0.404	0.322	†	†	†
LTCH	38	0.518	0.784	†	†	†
SNF	121	0.631	0.743	21	0.376	0.563
VI.C12 Injectable drug management	404	0.588	0.744	61	0.527	0.708
Acute	36	−0.080[a]	−0.142	†	†	†
HHA	94	0.785	0.880	19	0.830	0.855
IRF	115	0.447	0.465	†	†	†
LTCH	38	0.733	0.909	†	†	†
SNF	121	0.742	0.774	19	0.000	0.000
VI.C13 Make light meal	403	0.220	0.744	136	0.659	0.856
Acute	36	0.050	0.103	†	†	†
HHA	94	0.567	0.743	87	0.610	0.848
IRF	114	0.360	0.521	†	†	†
LTCH	38	0.582	0.742	†	†	†
SNF	121	0.713	0.732	31	0.795	0.937
VI.C14 Wipe down surface	404	0.594	0.765	153	0.653	0.805
Acute	36	0.064	0.123	†	†	†
HHA	94	0.572	0.685	87	0.620	0.799
IRF	115	0.432	0.636	22	0.401	0.791
LTCH	38	0.548	0.766	†	†	†
SNF	121	0.713	0.675	35	0.856	0.901
VI.C15 Light shopping	403	0.614	0.819	102	0.453	0.521
Acute	36	0.075	0.133	†	†	†
HHA	94	0.310	0.399	82	0.359	0.425
IRF	115	0.472	0.449	†	†	†
LTCH	38	0.621	0.771	†	†	†
SNF	120	0.785	0.766	†	†	†
VI.C16 Laundry	404	0.591	0.815	112	0.413	0.486
Acute	36	0.050	0.077	†	†	†
HHA	94	0.186	0.264	84	0.203	0.335
IRF	115	0.304	0.597	†	†	†
LTCH	38	0.625	0.798	†	†	†
SNF	121	0.787	0.807	†	†	†

(continued)

Table 2-11 (continued)
IRR testing: Function—supplemental self-care, mobility, and IADLs at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 6)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
VI.C17 Use public transportation	404	0.461	0.291	16	0.691	0.857
Acute	36	0.075	0.133	†	†	†
HHA	94	0.111	0.070	†	†	†
IRF	115	0.490	0.504	†	†	†
LTCH	38	0.374	0.449	†	†	†
SNF	121	0.700	0.503	†	†	†

NOTE: Correlations are reported for continuous items; kappas are reported unless otherwise noted. N/A: Weighted kappa is not applicable for items with only two response categories available. IRR sample: 456 pairs of assessments. HHA, home health agency; IADL, instrumental activities of daily living; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

[a] Kappa unable to be interpreted.

[b] Kappa unable to be interpreted. There were 13 agreements and 8 disagreements.

* With not attempted, not applicable, or refused excluded.

† Kappas for items with a sample size less than 15 are not reported.

SOURCE: RTI analysis of CARE data (data from CAREREL063 runs—LS20 [nstr] and LS10—data by provider type).

2.6 Section VII. Overall Plan of Care/Advance Care Directives

The CARE tool contains one item concerning a patient's overall health status and prognosis, which is designed to be a measure of patient frailty. A frail patient is likely to be readmitted to an acute hospital and have higher resource utilization. Although the OASIS-B assessment instrument contains a similar question evaluating a patient's risk of death within the next 6 months, the CARE item has been modified in that it includes a response category indicating "that a patient has serious progressive conditions that could lead to death within a year." It is based on the British Gold Standard.

This section of the tool also records the presence of agreed-upon care goals and whether care decisions have been documented in the patient's record. Results of the interrater reliability testing analysis for this section are shown in ***Table 2-12***. Overall kappas were substantial for these items. Patient's overall status had slightly lower, but still substantial, overall kappas.

For the "Agreed-upon care goals documented" item, kappas for LTCHs and SNFs showed almost perfect agreement and IRFs had substantial agreement. HHAs exhibited only fair agreement for this item, and the kappa for acute hospitals was unable to be interpreted. Provider-specific analysis of the overall status item (VII.A2) showed similar kappas across provider type. Agreement was generally lower on this item, with only moderate agreement in the unweighted

estimates except for respondents in LTCHs, which had substantial agreement. Estimates were higher with "unclear" and "unknown" responses excluded; however, they were still less than 0.60 except in LTCHs where kappas showed substantial agreement. For items VII.A3a and b ("Decision-maker designated" and "Decision to forgo resuscitation documented"), agreement was consistently substantial or higher, with almost perfect agreement in LTCHs for both questions and for HHAs on VII.A3a. Acute hospitals displayed moderate agreement on VII.A3a as did IRFs on VII.A3b.

Summary

This section describes the reliability of the overall plan of care items.

- Overall kappas were substantial or almost perfect for these items, and the patient's overall status item had slightly lower, but still substantial, overall kappas.

- The kappas for the provider-specific analyses were also generally substantial or almost perfect. HHAs exhibited fair agreement for the "Agreed-upon care goals documented" item, and the kappa for this item for acute hospitals could not be interpreted.

- The overall status item had lower provider-type specific kappas, with most showing only moderate agreement and only LTCHs showing substantial agreement.

Table 2-12
IRR testing: Patient overall status at PAC admission and acute discharge, IRR sample, by provider type (CARE Section 7)

Item	Effective sample size	Kappa	Weighted kappa	Effective sample size*	Kappa*	Weighted kappa*
VII.A1 Agreed-upon care goals documented	434	0.795	0.802	428	0.818	N/A
Acute	64	−0.016[a]	0.000	62	0.000	N/A
HHA	102	0.230	0.243	101	0.244	N/A
IRF	101	0.726	0.726	101	0.726	N/A
LTCH	49	0.799	0.812	48	0.832	N/A
SNF	118	0.887	0.895	116	0.921	N/A
VII.A2 Patient's overall status	434	0.617	0.765	410	0.592	0.680
Acute	64	0.545	0.666	64	0.545	0.666
HHA	102	0.554	0.726	102	0.554	0.726
IRF	101	0.615	0.813	80	0.425	0.353
LTCH	49	0.759	0.696	48	0.787	0.739
SNF	118	0.553	0.489	116	0.566	0.553
VII.A3a. Decision-maker designated	434	0.756	N/A	—	—	—
Acute	64	0.511	N/A	—	—	—
HHA	102	0.852	N/A	—	—	—
IRF	101	0.780	N/A	—	—	—
LTCH	49	0.870	N/A	—	—	—
SNF	118	0.750	N/A	—	—	—
VII.A3b. Decision to forgo resuscitation documented	434	0.786	N/A	—	—	—
Acute	64	0.797	N/A	—	—	—
HHA	102	0.784	N/A	—	—	—
IRF	101	0.564	N/A	—	—	—
LTCH	49	0.851	N/A	—	—	—
SNF	118	0.787	N/A	—	—	—

NOTE: N/A—Weighted kappa is not applicable for items with only two responses available. IRR sample: 456 pairs of assessments. HHA, home health agency; IRF, inpatient rehabilitation facility; IRR, interrater reliability; LTCH, long-term care hospital; PAC, post-acute care; SNF, skilled nursing facility.

[a] Kappa unable to be interpreted. There were 61 agreements and 3 disagreements.

*With unclear or unknown excluded.

SOURCE: RTI analysis of CARE data, IRR sample only (CARE extract 1/28/10).

[This page intentionally left blank.]

SECTION 3
SUMMARY AND CONCLUSIONS

This analysis builds on prior work by RTI examining the reliability of the standardized assessment items in the Continuity Assessment Record and Evaluation (CARE) Item Set for use in acute and post-acute settings. The prior work (documented in Volume 2 of *The Development and Testing of the Continuity Assessment Record and Evaluation [CARE] Item Set: Final Report on Reliability Testing*) included analysis looking at reliability estimates pooled across all acute hospital and post-acute setting types included in the Post-Acute Care Payment Reform Demonstration (PAC-PRD), with a focus on the performance of selected CARE items within provider types (inpatient rehabilitation facilities [IRFs], long-term care hospitals [LTCHs], home health agencies [HHAs], skilled nursing facilities [SNFs], and acute hospitals), but not all. The current report includes provider-type specific results for all CARE items tested in the interrater reliability analyses. An underlying goal of the CARE assessment design was to develop an assessment that captured patient clinical, functional, cognitive, and social support characteristics with equal or better reliability than comparable items on the assessment instruments currently mandated in post-acute settings. This additional stratification by provider type supplies additional important information on whether the performance of the CARE items, with regard to consistency of measurement, varies by the type of setting where the assessment is being completed.

Key findings are as follows:

- Reliability estimates for the vast majority of items evaluated were substantial or almost perfect, whether evaluating them across the full sample studied or when looking at reliability estimates within each provider setting.

- Although there were selected items with low agreement, measured by kappa, caution should be exercised in interpretation, as several of these were for items that measure conditions that occurred infrequently in the sample population (e.g., persistent vegetative state, tube feeding).

- With regard to provider-type specific analyses, most items performed well with substantial or better agreement statistics, though for some items variation in reliability across results was observed.

- For the sections on prior functioning (Section 2) and functional assessment (Section 6), IRFs had consistently lower kappa results than the other settings, though agreement was still moderate to substantial for most items. This difference in reliability may be reflective of greater variation in the functional status of IRF patients compared with patients receiving services from acute care hospitals, LTCHs, HHAs, and SNFs, rather than true differences in reliability across settings.

- The reliability of the skin integrity item assessing patient risk for developing pressure ulcers (III.G1) was only moderate to substantial (0.425–0.752) but was consistent across provider types. The lower kappa may be attributable to relatively few patients being assessed as at risk.

- Items in the cognitive status, mood, and pain section (Section 4) performed very well in each provider type, with few exceptions. The sample size of patients receiving the observational assessment of cognitive status (IV.C) was too small to evaluate the performance of those items, and the lower number of patients administered the Confusion Assessment Method (IV.D) lowered the kappas for a few items included in that inventory. No one setting had systematically lower reliability for any of these sets of items.

Results from these analyses show that CARE items are consistently reliable when examined overall and also when examined within and across acute and post-acute provider-type settings. Results show that the CARE tool can be implemented in all of the settings studied, yielding patient assessment results that are at least as reliable as those produced by the previously used patient assessment tools. The kappa results for CARE items are equivalent to or better than reported agreement statistics available from testing of items capturing similar concepts on the Minimum Data Set, Outcome and Assessment Information Set, and FIM®, where comparable results were available. This evidence supports the assertion that the standardized set of items included in the CARE Item Set can be used to measure Medicare beneficiaries' health status in a consistent way across acute hospital, LTCH, IRF, SNF, and HHA settings.

REFERENCES

ABT Associates: Validation of Long Term and Post-Acute Care Quality Indicators. Prepared for the Office of Clinical Standards and Quality: Centers for Medicare and Medicaid Services. Contract Number 500-95-0062/Task Order #4, 2003.

Berg, K.: Appendix G: OASIS+ Interrater Reliability Study. In: ABT Associates Inc.: Case-Mix Adjustment for a National Home Health Prospective Payment System: Second Interim Report. Prepared for Office of Strategic Planning: Health Care Financing Administration. Contract Number 500-96-0003/Task Order #2, 1999.

Gage, B., Morley, M., Constantine, R., et al.: Examining Relationships in an Integrated Hospital System. Contract No 06EASPE060059. Waltham, MA: RTI International, March 2008.

Gage, B., Smith, L., Ross, J., et al.: The Development and Testing of the Continuity Assessment Record and Evaluation (CARE) Item Set: Final Report on Reliability Testing: Volume 2. Contract HHSM-500-2005-00291. Research Triangle Park, NC: RTI International, August 2012.

Hamilton, B., Laughlin, J., Fiedler, R., et al.: Interrater reliability of the 7-level functional independence measure (FIM®). Scandinavian Journal of Rehabilitation Medicine 26(3):115-119, 1994.

Hirdes, J.P., Smith, T.F., Rabinowitz, T., et al.: The Resident Assessment Instrument-Mental Health (RAI-MH): Interrater reliability and convergent validity. Journal of Behavioral Health Services and Research 29(4):419-432, 2002.

Hittle, D., Shaughnessy, P., and Crisler, K.: A study of reliability and burden of home health assessment using OASIS. Home Health Care Services Quarterly 22(4):43-63, 2002.

Iowa Foundation of Medical Care: Staff Time and Resource Intensity Validation. Interrater Reliability Worksheet. West Des Moines, Iowa, n.d. Data retrieved September 19, 2007.

Madigan, E., and Fortinsky, R.: Interrater reliability of the Outcomes Assessment Information Set: Results from the field. The Gerontologist 44(5):689-692, 2004.

Mor, V., Angelelli, J., Jones, R., et al.: Interrater reliability of nursing home quality indicators in the U.S. BMC Health Services Research 3(20):1-13, 2003.

Morris, J., Nonemaker, S., Murphy, K., et al.: A commitment to change: Revision of HCFA's RAI. Journal of the American Geriatrics Society 45(8):1101-1106, 1997.

Nagi, S.Z.: Some conceptual issues in disability and rehabilitation. In M.B. Sussman (Ed.), Sociology and Rehabilitation (pp. 100-113). Washington, DC: U.S. Department of Health, Education, and Welfare, 1965.

RAND Health Corporation: Development and Validation of a Revised Nursing Home Assessment Tool: MDS 3.0. Prepared for the Office of Clinical Standards and Quality: Centers for Medicare and Medicaid Services. Contract Number 500-00-0027/Task Order #2, 2008.

Streiner, D.L., & Norman, G.R.: Health Measurement Scales: A Practical Guide to Their Development and Use (Second ed.). Oxford: Oxford University Press, 1995.