September 2012

Continuity Assessment Record and Evaluation (CARE) Item Set: Video Reliability Testing

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
Daniel Barch, MS
Jessica Ross, MPH
Kate Shamsuddin, BS
Judith Hazard Abbate, PhD
Carole Schwartz, MS, OTR/L
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:

Video Reliability Testing

by Laura Smith, PhD
Anne Deutsch, RN, PhD, CRRN
Daniel Barch, MS
Jessica Ross, MPH
Kate Shamsuddin, BS
Judith Hazard Abbate, PhD
Carole Schwartz, MS, OTR/L
Barbara Gage, PhD

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. 500-2005-00291. 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

Video “Patients”: Phillip, Deb, Octavia, Kate, Joe, Mr. Jones,
Dorian, Ms. Smith, and John

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

Thank you to our many colleagues across the RTI team. Our colleagues provided invaluable assistance with the development of the video “patients,” training, and other important inputs needed to complete this data collection successfully. These colleagues include

The Rehabilitation Institute of Chicago, especially Hsiang-Yi Tseng, Linda Ehrlich-Jones, Caitlyn Taylor, and Kyle Seanor; and

The Visiting Nurse Service of New York, especially Sara Couch, Ann-Marie Kamuf, and Shivani Shah.

[This page intentionally left blank.]

CONTENTS

List of Tables

EXECUTIVE SUMMARY

E.1 Introduction

E.1.1 Background

The Continuity Assessment Record and Evaluation (CARE) Item Set was developed as part of the larger Post-Acute Care Payment Reform Demonstration (PAC-PRD), authorized by the Deficit Reduction Act of 2005. It was developed as a standardized set of items for measuring medical, functional, cognitive, and social support factors in the acute hospital, long-term care hospital (LTCH), inpatient rehabilitation facility (IRF), skilled nursing facility (SNF), and home health agency (HHA) settings to provide a way to compare the health status of Medicare beneficiaries across provider types. The Item Set was also designed to elicit consistent responses from caregivers of different disciplines. Responses on CARE items after viewing videos of standard patients generally showed good agreement: respondents tended to agree with each other and with a team of clinical experts that previewed and rated the videos. In our comparisons of rates of agreement by provider type, respondents from SNFs tended to have the highest rates of agreement with the sample of participants and the clinical expert team. Across a wide array of items, when respondents disagreed with the clinical expert team, they tended to provide responses that indicated better health status than did the responses of the experts; however, this pattern is consistent across provider types and clinician discipline. On functional items, agreement rates tended to be higher when evaluating high-functioning patients than when evaluating low-functioning patients, a finding that echoes similar research from earlier studies of similar functional assessment tools (Fricke, Unsworth, and Worrell, 1993).

E.1.2 Overview

Standardization of patient assessment data collected across acute and post-acute care (PAC) settings was a primary goal of the development of the CARE Item Set; therefore, consistency of ratings, or reliability, of CARE items across these provider settings is crucial. A well-designed item set for evaluating patients across the continuum of acute and PAC settings should, given the same set of patients to evaluate, elicit similar responses across all provider types and credentials of clinicians performing assessments. To that end, videos of patients were developed and distributed to a subset of acute and PAC providers participating in the PAC-PRD data collection. These providers supplied staff of various disciplines to perform assessments, so that a wide range of clinician participants would be able to evaluate the same set of patients. Using standard video patients made it possible to examine analytically whether clinicians with different credentials practicing in different provider settings agreed in their assessment of the same patients. This report expands on results from prior analyses 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). In this report, RTI examined the impact of clinician discipline, clinician familiarity with specific sections of the CARE Item Set (i.e., was the clinician assigned to fill out similar items in the larger PAC-PRD data collection?), and provider type. In addition, RTI examined whether the health status of patients had a potential influence on item reliability by comparing rates of agreement when assessors were rating patients with high levels of function and patients with low levels of function. Finally, RTI examined cases in which participating clinicians disagreed with the clinical expert team to see whether those clinicians with different credentials or practicing in different settings systematically rated patients as having better or worse health status.

E.2 Methods

Nine videos, each approximately 20 minutes in duration and depicting an aging patient, were created by the Rehabilitation Institute of Chicago (RIC). The patients varied in their primary diagnosis, skin integrity, cognitive impairments, functional ability, and modes of mobility. A clinical expert team, comprising clinicians from RIC, RTI, and the Visiting Nurse Service of New York (VNS-NY) viewed these videos, evaluated the patients using the CARE Item Set, and made clarifying alterations to the video. Clinicians were then recruited from 28 facilities that were already taking part in the PAC-PRD and were tasked with evaluating the patients depicted in the videos using the CARE Item Set. The video reliability testing included analyses of rates of agreement between participating clinicians and the modal response of the set of participating respondents and with the responses of the RTI clinical expert team.

E.3 Results

E.3.1 Sample: Assessor Demographics

To better understand the characteristics of our group of clinicians, we calculated frequencies and proportions by the discipline of responding clinicians and the type of provider in which those clinicians worked. Registered nurses (RNs) completed the largest proportion of assessments (47 percent), followed by physical therapists (PTs; 21 percent), occupational therapists (OTs; 14 percent), and “other” (8 percent). IRFs contributed the most assessments of any provider type (43 percent), followed by HHAs (21 percent), SNFs (12 percent), and acute hospitals (3 percent).

E.3.2 Impact of Clinician Discipline and Familiarity on Agreement by Item

For the CARE Item Set to be reliable, identical patients should ideally be assessed with identical responses regardless of the discipline of the clinician completing the assessment. Further, after completing training, all clinicians should respond with identical responses, regardless of whether those caregivers typically assess certain aspects of health status in the course of their work (e.g., PTs and functional assessment items). To help evaluate whether items elicited consistent responses from clinicians regardless of their typical area of practice and prior familiarity with the CARE items, respondents were also asked which aspects of health status they typically assessed using CARE during the larger PAC-PRD data collection. Results showed that RNs, PTs, and OTs had similar rates of agreement with the clinical expert team, whereas the “other” group was significantly different. There was no significant difference between clinicians who responded “yes” that they typically assessed an item and those who responded “no” that they did not typically assess an item.

E.3.3 Impact of Provider Type on Rates of Agreement

A reliable assessment should elicit consistent responses from clinicians even if they are practicing in different provider settings. For example, for the same set of patients, clinicians working in acute-care hospitals should show similar patterns of responses to clinicians working in SNFs (holding patient characteristics constant). Provider type should have little to no impact on consistency of ratings made by clinicians of the same discipline upon viewing the video patients. An analysis showed that rates of agreement varied little between different provider types, although this difference was consistent enough to produce a statistically significant effect.

E.3.4 Rates of Agreement among Registered Nurses in Different Provider Types

Clinicians of the same discipline practicing in different provider settings should rate the same patient the same way, regardless of the provider setting where the assessment was completed. Because sufficiently large numbers of RNs responded in each facility type, those responses were analyzed to see if there were any differences attributable to facility type while holding discipline constant. There was little difference across the rates of agreement by provider type within the set of registered nurses who participated in the data collection.

E.3.5 Impact of Provider Type and Functional Dependency of Patient on Rates of Agreement

To further examine the reliability of the CARE Item Set across provider type and clinician discipline, RTI examined whether rates of agreement on CARE items were similar for patients with differing health status. Clinician agreement was examined for the items in the Functional Status section (VI) of CARE, focusing on patients with a high level of function as compared to patients with a low level of functioning. Rates of agreement tended to be higher for high functioning status patients than for patients with low functional status. This trend held across all care settings. However, these results mirrored those found in prior research examining functional assessment using a different tool.

E.3.6 Impact of Provider Type, Clinician Discipline, and Patient Functional Status on Direction of Disagreement

Disagreements with the clinical expert team were analyzed to check for patterns of responses—that is, whether responding clinicians tended to be more or less optimistic regarding the health status of the videotaped patients. Compared with the clinical expert team, responding clinicians tended to give ratings that indicated generally more positive health status. This trend was present for all settings but was significant only when pooling all respondents (after controlling the type I error rate for multiple comparisons).

E.4 Conclusions

Video testing indicated that the CARE Item Set is reliable across provider types, and, with training on the proper completion of the CARE assessment, clinicians can produce consistent responses regardless of discipline. Results were similar to other patient assessments when considering consistency of responses in evaluating patients with higher or lower functional capacity, but did indicate that responses were more consistent when clinicians were rating patients with higher levels of independence. Results also showed that respondents tended to assess patients as having better health status than did the team of clinical experts; however, this finding was consistent across assessor discipline and provider type. In sum, results from the analyses described in this report support the assertion that the CARE could be used effectively to obtain consistent patient assessment across acute and PAC provider types.

[This page intentionally left blank.]

SECTION 1
INTRODUCTION

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). While 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 to be 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 behind the currently mandated assessment items in the Medicare payment systems, including those in the mandated Inpatient Rehabilitation Facility Patient Assessment Instrument (IRF-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 comment. 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). More than 40,000 assessments were collected in acute hospitals, LTCHs, IRFs, SNFs, and HHAs. An additional 455 assessments were collected as part of a test of item reliability.

1.2 Purpose

Video reliability tests were designed to measure the level of clinician agreement across levels of care. A range of clinicians in each provider type were asked to assess a standard set of patients presented through a videotape of a patient evaluation. This process ensured that the same information was presented to each clinician and allowed examination of differences in scoring among different types of clinicians examining the same patient.

The goal of the CARE item development is to standardize items used across multiple health care provider types. Therefore, it is important that CARE items consistently capture variation in patient health status both within and across populations. Given the variety of clinical disciplines that may be providing services across the continuum of PAC providers, this report summarizes findings from analyses examining whether an assessor's disciplinary background (e.g., registered nurse, physical therapist, occupational therapist) or setting of care (e.g., acute hospital, LTCH, IRF, SNF, and HHA) impacts the ability to consistently measure a patient's health status. We evaluate this question by analyzing the ability of clinicians from varying disciplines and provider types to assess a standard set of nine patients presented via video using the CARE items. This report adds to the work 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), which reported on consistency of rating among clinicians on CARE items, looking at strict agreement with the modal response for each item for that "patient" and with the response selected by the expert clinicians who designed the videos. This report reflects expanded analyses, examining whether clinician training is associated with assigning systematically higher or lower ratings for patients and whether provider type and patient severity affect rates of agreement. Analyses also examine the performance of items for raters who ordinarily filled out that item in the larger PAC-PRD data collection and raters who did not usually fill out that item.

SECTION 2
METHODS

2.1 Video Criteria and Development

The videos for this part of the reliability testing were developed by key RTI project staff, clinicians, and subcontractors, with input from CMS. The team developed a total of nine videos to distribute to the providers participating in video testing. The patient "case studies" in each of the videos vary by medical complexity, functional abilities, and cognitive impairments. The nine videos allowed patients to be classified as high, medium, or low on each of these three factors. Each facility or agency received three videos where at least one video demonstrated the following elements: cognitive impairments, skin integrity problems, a wheelchair-dependent case study patient, and a variety of midlevel functional items. The midlevel functional items were considered to be the most challenging for clinicians to score and are thus of particular interest in establishing reliability. **Table 2-1** provides a brief description of the clinical characteristics of each of the nine video "patients."

The Rehabilitation Institute of Chicago (RIC), a subcontractor on the CARE Item Set development project, created, revised, and edited the nine videos for testing use. Each video underwent two phases of review. First, the reliability team internally reviewed the videos through a multistep process. This process began with the range of clinicians from RTI, RIC, and the Visiting Nurse Service of New York (VNS-NY) watching the videos, scoring the corresponding tools, and submitting responses anonymously. Once the scores were compiled, the clinicians met to discuss the content of the videos as well as any discrepancies in scoring; at least five clinicians with various clinical backgrounds (nursing, rehabilitation, and home health) attended each of the video review meetings. The clinicians agreed to and submitted clarifying revisions and edits for each of the videos. These revisions commonly consisted of clarifying voiceovers. The clinical team repeated the process of viewing, scoring, discussing, editing, and finalizing the videos until all nine were ready for distribution. This method allowed us to reach consensus by the internal clinical team on the scoring of each item.

This work provides valuable insight on whether the CARE items can be used reliably by clinicians of diverse clinical backgrounds and provider types. In addition, because there is relatively high turnover of staff in health care settings, the ability of a relatively brief training to produce acceptably consistent ratings is important.

2.2 Sample Selection and Data Collection

RTI estimated the required sample size for this work and determined that approximately 5–10 unique providers should be recruited from each of the five levels of care (acute-care hospitals, HHAs, IRFs, LTCHs, and SNFs). Participants in this part of the data collection were selected from the nearly 150 providers within the PAC-PRD market areas, focusing particularly on providers that were midway through their CARE data collection; many of the same providers that participated in the interrater reliability tests participated in this component. RTI recruited 28 providers from the set of providers already enrolled in the PAC-PRD data collection. See **Table 2-2** for counts of providers and the number of assessments submitted by provider type.

All CARE-trained clinicians from acute hospitals, LTCHs, IRFs, SNFs, and HHAs participating in the interrater reliability testing were asked to watch three short videos and assess patient "case studies." Only staff previously collecting CARE information in the demonstration participated in video reliability testing. All assessors collecting data for PAC-PRD were licensed professionals. Overall, nurses almost always completed the CARE medical items, but the impairments and functional items were completed by nurses, physical therapists, occupational therapists, and, when appropriate, speech pathologists. The cognitive section was completed by nurses, occupational therapists, speech therapists, and case managers, depending on the individual facility. Case managers or discharge planners frequently completed the sections on Overall Plan of Care/Advanced Directives and Discharge Status. Some organizations chose to physically divide the tool by discipline for completion. Other organizations used different staff to complete different sections of the same paper tool. For example, physical therapists and occupational therapists often divided the Functional Status section of the tool. A small number incorporated CARE into their reporting systems. Some organizations, such as HHAs, used one assessor for each patient, usually the nurse, unless the admitting discipline was a physical therapist, in which case, they completed the items.

Each demonstration site identified the clinician(s) who would participate in the video reliability data collection. To account for different lengths of time that had elapsed since the initial PAC-PRD CARE training in each market, each clinician participating in the video testing attended a 1.5-hour CARE refresher training before beginning the data collection. After the CARE refresher trainings, RTI also reviewed the video data collection instructions with the demonstration project coordinators. Each clinician involved in reliability testing was asked to view three short videos and assess these patient "case studies" in accordance with the guidelines and protocols developed by RTI. Each video was approximately 20 minutes in length and had a corresponding CARE Item Set, with the items arranged in the sequence in which they appeared in the respective video.

During the video portion of the reliability testing, RTI instructed each staff member to fill out the entire CARE Item Set despite ordinary practices for data collection. Clinicians were instructed to document, in advance of scoring the "case studies," their typical practices for completing the CARE Item Set. The collected information had two main components: (1) whether the clinician attended the CARE Item Set refresher session, and (2) which subsections of the CARE Item Set he or she usually completed or did not complete. Clinicians were instructed to code what they saw and heard as each activity was presented even if clinical experience indicated otherwise. Additionally, clinicians were asked to use independent judgment when scoring a patient's status and not discuss CARE item scores with other clinicians until all participating clinicians had submitted completed CARE Item Set forms to the project coordinator or backup coordinator.

Table 2-1
Patient case study characteristics by video

Video	Phillip (1)	Octavia (2)	Kate (3)	Joe (4)	Mr. Jones (5)	Deb (6)	Dorian (7)	Ms. Smith (8)	John (9)
Diagnosis	Parkinson's disease	Cerebral vascular accident	Chronic obstructive pulmonary disease exacerbation	Total knee arthroplasty	Mild myocardial infarction deconditioning	Shoulder surgery	Fall with injury to stump	Hip fracture	Closed head injury Knee surgery
Skin integrity	Pressure ulcer	Intact	Intact	Intact	Intact	Pressure ulcer	Intact	Intact	Pressure ulcer
Cognitive impairments	No	Yes	No	No	Yes	Yes	No	Yes	Yes
Functional ability	Low	Medium	High	High	Medium	Low	High	Medium	Low
Mode of mobility	Walks	Wheels	Walks	Walks	Walks	Wheels	Wheels	Wheels	Walks

Table 2-2
Video testing providers by type/level of care

Provider type	Number of providers enrolled	Video assessment numbers
Acute hospitals	3	15 assessments
Home health agencies (HHAs)	9	118 assessments
Inpatient rehabilitation facilities (IRFs)	8	237 assessments
Long-term care hospitals (LTCHs)	3	114 assessments
Skilled nursing facilities (SNFs)	5	66 assessments
Total	**28**	**550 assessments**

RTI initially conducted a small pilot in the Boston market area to test and refine the video reliability testing materials, including the videos, tools, and instructions. At the time of the pilot, the participating clinicians held positions in facilities or agencies across four levels of care. The pilot viewers were nurses, physical therapists, or occupational therapists by background. Any of the clinicians from the participating sites who viewed the pilot videos were excluded from participation in the subsequent, full reliability video testing. CMS staff also participated in the reviews. The pilot viewers provided comments and suggestions on several aspects of the videos. On the basis of this feedback, further revisions were made prior to the full reliability video testing.

2.3 Item Selection for Testing

CARE Item Set items selected for video testing fell into one (or more) of the following categories: items that were 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.

2.4 Analyses

Multiple analytic approaches were used for assessing the video reliability of the CARE Item Set items, adhering to, and building on, the methods used by Fricke and colleagues to assess the reliability of the FIM®[1] items using videos (Fricke, Unsworth, and Worrell, 1993). First, for each CARE item included in at least one of the nine videos, percent agreement with the modal response was calculated. In initial analyses, RTI did not consider agreement at one response level above and below the mode but instead used a stricter approach, looking at direct modal agreement only. In the second approach, percent agreement with the internal clinical team's

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

consensus response was also calculated. This second measure not only gives an indication of item reliability but reflects on training consistency.

For the analyses included in this report, agreement was analyzed for the whole sample of respondents as a function of respondent clinician discipline, familiarity with specific sections of the CARE Item Set, and provider type. In addition, to examine possible differences in the evaluation of patients of varying health status, RTI examined respondent agreement for the three low-functioning patients in our videos and for the three high-functioning patients. Finally, for responses that differed from the responses of the clinical expert team, RTI analyzed the direction of discrepancy: that is, whether those discordant responses indicated better or worse health status than did the evaluation represented by the clinical experts. We stratified these analyses by discipline and provider type to examine whether there were any systematic differences in the direction of disagreement across provider types and across rater discipline.

[This page intentionally left blank.]

SECTION 3
RESULTS

Rates of agreement among the responding clinicians—and between the responding clinicians and our team of clinical experts—were generally high for the vast majority of items evaluated. For the remaining items, low sample sizes and lack of familiarity with specific kinds of assessments (e.g., speech therapists do not frequently assess functional ability) depressed agreement rates. The first round of results, reported on in greater detail in a prior report, indicated that the CARE Item Set can be used effectively across provider types and by clinicians of various disciplines (Gage et al., 2012). These results were also consistent with results reported for the more traditional paired interrater reliability testing summarized in the same report.

On the basis of these results, RTI performed further analyses to more closely examine factors that may have influenced the level of agreement on CARE Item Set responses: the familiarity of clinicians with certain assessment items through their ongoing participation in the PAC-PRD data collection, the respective disciplines of the clinicians, the provider type in which the clinicians worked, and the functional status of the patient being assessed.

Generally, clinicians showed high rates of agreement in assessing the patients portrayed in the videos, agreeing both with each other and with the clinical expert team. Clinicians of most disciplines had comparable rates of agreement, as did clinicians both familiar and unfamiliar with the different clinical domains and subsections of the CARE assessment. The health care setting where respondents worked had a weak effect on their level of agreement, although this effect was not seen after holding respondent discipline constant (i.e., looking only at registered nurses in various settings). Analysis of patterns of disagreement reveals a pair of trends that may inform future development of and training for the CARE Item Set. First, rates of agreement tended to be higher for high-functioning patients than for low-functioning patients on items that assessed functional ability. It should be noted that this finding is consistent with analyses of other functional assessment instruments (Fricke et al., 1993). Also, there was a general tendency for clinicians to indicate better health status for the video patients than did the clinical expert team, but this tendency was consistent across provider types. These are important trends to monitor and address in future use of the CARE Item Set. However, this set of analyses indicates that reliability for the CARE Item Set is generally good for clinicians of different disciplines, experiences, and provider types.

3.1 Sample: Assessor Demographics

In this section, RTI describes the distribution of clinician disciplines and provider types represented among the assessments received during the video reliability data collection period.

Summary of Findings

- Registered nurses were the best-represented discipline of all groups of respondents, comprising 47.1 percent of the sample.

- IRFs contributed 43 percent of responses.
- Acute facilities contributed just 3 percent of responses.

Detailed Findings

Tables 3-1a and **Table 3-1b** show the basic characteristics of the clinicians who assessed the videos, in terms of both their discipline and provider type.

Table 3-1a indicates that the highest proportion of assessments was completed by registered nurses (RNs), at 47 percent, followed by physical therapists (PTs) at 21 percent and occupational therapists (OTs) at 14 percent. The category of "other," which is comprised mostly of licensed nurse practitioners (LPNs), made up 8 percent of the assessments. Case managers and speech therapists contributed 6 percent and 5 percent, respectively. **Table 3-1b** shows that IRFs contributed the most video assessments (43 percent), followed by HHAs (22 percent), LTCHs (21 percent), SNFs (12 percent), and acute providers (3 percent).

3.2 Impact of Clinician Discipline and Familiarity on Agreement by Item

For the CARE Item Set to be reliable, the ratings of caregivers of various disciplines should agree in their assessments of identical patients (as facilitated by the use of videotaped patients). A well-designed assessment should obtain consistent results from caregivers regardless of whether they typically assess certain aspects of health status in the course of their work, given that they have completed training on the CARE Item Set. Therefore, in addition to identifying their discipline, respondents were asked which segments of the CARE assessment they typically completed.

Early testing indicated, as hypothesized, that agreement on CARE items was lower among groups of clinicians who do not typically encounter those items in their regular practice (e.g., speech language pathologists and items assessing skin integrity). Agreement rates were analyzed for the effect of clinician discipline and familiarity with the CARE items they were completing. The effect of clinician discipline was tested overall for the full set of items, and then in pairwise comparisons of different clinician disciplines (e.g., RN vs. PT). These comparisons were meant to identify which disciplines had similar patterns of agreement and which settings were more distinct from each other.

Summary of Findings

- Rates of agreement were similar for most disciplines.
- Results of pairwise comparisons between disciplines showed that the difference in rate of agreement among RNs and among OTs was small (5.0 percent) but significant.
- The "other" group was significantly different in their rate of agreement than each specific discipline.

Table 3-1a
Clinicians completing video assessments by discipline

Clinician type	Phillip (1)	Octavia (2)	Kate (3)	Joe (4)	Mr. Jones (5)	Deb (6)	Dorian (7)	Ms. Smith (8)	John (9)	Total
Case mgr (n/%)	3	3	3	3	3	3	5	5	5	33
Case mgr (n/%)	4%	5%	4%	7%	7%	7%	7%	7%	7%	6%
OT (n/%)	10	4	9	7	7	7	10	10	10	74
OT (n/%)	13%	7%	13%	16%	16%	16%	14%	14%	14%	14%
PT (n/%)	16	9	16	9	9	9	16	15	15	114
PT (n/%)	21%	15%	23%	21%	21%	21%	22%	21%	21%	21%
RN (n/%)	29	27	25	22	22	22	37	38	37	259
RN (n/%)	39%	45%	35%	51%	51%	51%	51%	53%	52%	47%
Speech (n/%)	4	4	4	0	0	0	4	4	4	24
Speech (n/%)	5%	7%	6%	0%	0%	0%	6%	6%	6%	5%
Other (n/%)	13	13	14	2	2	2	0	0	0	45
Other (n/%)	17%	22%	20%	5%	5%	5%	0%	0%	0%	8%
Total (n/%)	75	60	71	43	43	43	72	72	71	550
Total (n/%)	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

NOTE: Percent = column percent, Case mgr = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, Speech = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: JR18).

Table 3-1b
Clinicians completing video assessments by provider type

Clinician type	Acute	LTCH	IRF	SNF	HHA	Total
Case mgr (n/%)	3	6	21	3	0	33
Case mgr (n/%)	9%	18%	64%	9%	0%	100%
OT (n/%)	0	12	50	12	0	74
OT (n/%)	0%	16%	68%	16%	0%	100%
PT (n/%)	0	21	65	6	22	114
PT (n/%)	0%	18%	57%	5%	19%	100%
RN (n/%)	12	48	82	21	96	259
RN (n/%)	5%	19%	32%	8%	37%	100%
Speech (n/%)	0	9	15	0	0	24
Speech (n/%)	0%	38%	63%	0%	0%	100%
Other (n/%)	0	18	4	24	0	46
Other (n/%)	0%	40%	9%	52%	0%	100%
Total (n/%)	15	114	237	66	118	550
Total (n/%)	3%	21%	43%	12%	22%	100%

NOTE: LTCH = long-term care hospital, IRF = inpatient rehabilitation facility, SNF = skilled nursing facility, HHA = home health agency, Percent = row percent, Case mgr = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, Speech = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: JR18).

- There was no significant difference in rates of agreement between respondents who indicated that they typically assessed a given aspect of health status and those who indicated that they did not.

Detailed Findings

The effect of clinician discipline and familiarity with the CARE items being completed was examined overall using a two-way repeated-measures general linear model to test the influences of clinicians' discipline and familiarity with the CARE items they were completing, as well as to test for an interaction between the two factors. **Table 3-2a** presents the results of this analysis.

Table 3-2a
General linear model for respondent discipline and familiarity on agreement by item

Source	Df	Sum of squares	Mean squares	*F*	*P*
Discipline	5	5.11	1.02	23.90	<0.0001
Familiarity	2	1.00	0.50	11.66	<0.0001
Discipline x Familiarity	8	10.39	1.30	30.35	<0.0001
Error	1,697	72.61	0.043		
Total	1,712	89.11			

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB03).

The analysis showed statistical evidence of an effect of discipline and item familiarity and an interaction between the two on rates of agreement. To examine whether there were any specific clinician disciplines that were driving the overall findings shown in Table 3-2a, we conducted further analyses. Comparing mean agreement rates between each pair of disciplines and each pair of item familiarity responses indicated that these effects may be driven by lower agreement among (1) respondents with "other" discipline and (2) respondents who have missing responses in the field meant to indicate that they typically fill out a section of the care tool. In fact, there was no significant difference between the mean rate of agreement between clinicians who indicated that they did typically fill out given sections of the item set and those who indicated that they did not. However, there were significant differences between these groups and those clinicians who did not provide information on their familiarity with an item. These comparisons are shown in **Table 3-2b**. Similarly, with regard to discipline, clinicians with the discipline categorized as "other" had rates significantly different than those of every other group. There was only one significant difference between groups with specific disciplines: the mean rate of agreement among RNs was 5 percentage points lower than the rate of agreement among OTs, a small disparity that nonetheless is statistically significant. It appears likely that the effect of discipline on rate of agreement was largely driven by much higher rate of agreement within the "other" group. The "other" group was largely made up of LPNs. **Table 3-2c** shows the results of these pair-wise comparisons between respondents by discipline.

Table 3-2b
Differences in mean agreement by familiarity with CARE section

Response to "usually assess"	Yes	No	Missing
Yes	—	−1.3%	−6.3%*
No		—	−5.0%*
Missing			—

**p* < .05, adjusted for multiple comparisons

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB03).

Table 3-2c
Differences in mean agreement by item between disciplines

Discipline	Case mgr	OT	PT	RN	ST	Other
Case mgr	—	1.1%	−0.9%	−3.9%	−0.7%	−14.4%*
OT		—	−2.0%	−5.0%*	−1.8%	−15.5%*
PT			—	−3.1%	0.2%	−13.5%*
RN				—	3.2%	−10.5%*
ST					—	−13.7%*
Other						—

**p* < .05, adjusted for multiple comparisons.

NOTE: Case mgr = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, ST = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB03).

3.3 Impact of Provider Type on Rates of Agreement

When using a reliable standardized assessment tool, clinicians from different provider types should show similar patterns of agreement in assessments of identical patients. Thus, when comparing mean rates of agreement across different provider types, there should not be wide variation. However, it is possible that experiences and practices that are specific to different provider types may influence the way clinicians assess a patient; for example, the case mix seen in an IRF setting might lead a clinician to assess a patient as more or less functionally dependent than a clinician who routinely works with the mix of patients usually seen in a SNF.

This pair of analyses examined whether the type of provider in which a clinician is assessing patients had an effect on rates of agreement. First, to see if clinicians working for different provider types showed different patterns of agreement, mean rates of agreement were examined. This analysis examined a possible influence of provider type on interpretation of the item set. Second, mean agreement rates were compared for each possible pair of provider types (e.g., comparing the differences in mean agreement rate between clinicians working in acute hospitals and between clinicians working in SNFs). This follow-up analysis examines the specific differences, if any, in patterns of agreement between clinicians of different provider types.

Summary of Findings

- The type of provider had a weak effect on rates of agreement when evaluated overall. However, when evaluated in pairwise analyses, the only significant difference in rates of agreement identified between specific facility types was the small difference (6.0 percent) between SNFs and IRFs.

Detailed Findings

The general linear model constructed to analyze rates of agreement as a function of provider type alone is summarized in **Table 3-3a**. The model indicates statistical evidence for an effect of provider type on rates of agreement ($F = 2.79$, $p < .05$), suggesting that the practices, resources, and case mixture at each provider type had some bearing on how clinicians assessed the videotaped patients using the CARE Item Set. However, this effect is extremely weak: provider type accounts for only 1.3 percent of the variance in rates of agreement. Thus, the role that provider type plays in the way respondents assess patients using the CARE Item Set is likely too small to be substantial.

A closer examination of the differences in mean agreement using pairwise comparisons shows that SNFs have higher mean rates of agreement than any other type of facility. However, a Tukey Honestly Significant Difference test, which controls the overall type I error rate at 0.05, showed that the only significant difference between facility types was the difference of 6 percentage points between SNFs and IRFs. The mean agreement rates of acute hospitals and LTCHs were nearly identical, and rates for both of these provider types differed from that of IRFs by less than 1 percent. Again, there were only small differences in rates of agreement between different care-giving settings. **Table 3-3b** displays the differences between the means of each pair of provider type.

Table 3-3a
General linear model for rates of agreement by provider type

Source	df	Sum of squares	Mean squares	*F*	*p*
Provider type	4	0.28261761	0.07065440	2.79	0.0259
Error	561	14.22120638	0.02534974		
Total	565	14.50382399			

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB06R).

Table 3-3b
Differences in mean agreement by item between provider types

Provider type	Acute	HHA	LTCH	IRF	SNF
Acute	—	−2.8%	0.0%	0.9%	−5.1%
HHA		—	2.8%	3.6%	−2.3%
LTCH			—	0.8%	−5.1%
IRF				—	−6.0%*
SNF					—

**p* < .05, adjusted for multiple comparisons.

NOTE: Acute = acute care facility, HHA = home health agency, LTCH = long-term care hospital, IRF = inpatient rehabilitation facility, SNF = skilled nursing facility.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB06R).

3.4 Rates of Agreement among Registered Nurses in Different Provider Types

Having examined rates of agreement in different settings, the next step was to examine patterns of agreement in a discipline-specific manner. In this section, we describe an analysis done to examine whether clinicians with the same discipline, but practicing in different provider settings, were similar in their ratings on CARE items. The characteristics of a given provider type may have a larger influence on patterns of agreement for specific disciplines than when all clinicians are examined together. This analysis holds discipline constant and looks for differences in mean rates of agreement between facility types. Because most disciplines represented among participating providers had very small numbers in certain provider types, and others had none at all (see **Table 3-1b**), we focused our analyses on RNs, who were represented across all of the participating setting types.

Summary of Findings

- The mean rate of agreement among RNs was 86.3 percent.
- When looking at the potential effect of the provider type in which different responding RNs were practicing, variation from the mean was small, ranging from 2.4 percentage points above (SNFs) to 1.8 percentage points below (HHAs).

Detailed Findings

Because not all RNs filled out each item, weighted means were taken to determine overall agreement rates for the different care provider types. **Table 3-4** presents these weighted means. Rates of agreement varied little between provider types, ranging from 84.5 percent for HHAs to 88.8 percent for SNFs. In fact, three of the provider types (acute, IRF, and SNF) had rates of agreement for RNs that fell within four-tenths of one percentage point.

Table 3-4
Weighted mean rates of agreement for registered nurses by provider type

Acute (12 total respondents)	HHA (96 total respondents)	IRF (82 total respondents)	LTCH (48 total respondents)	SNF (21 total respondents)	Total (259 total respondents)
88.6%	84.5%	88.2%	84.8%	88.8%	86.3%

NOTE: Acute = acute care facility, HHA = home health agency, IRF = inpatient rehabilitation facility, LTCH = long-term care hospital, SNF = skilled nursing facility. Because not all respondents assessed each item, means are weighted by the number of respondents per facility and per item.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB06).

3.5 Impact of Provider Type and Functional Dependency of Patient on Rates of Agreement

Another important question to address is whether certain patient characteristics lead to different patterns of responses on the CARE Item Set. Certain patient characteristics may lead to more frequent discrepant ratings: for example, patients who are healthier and less impaired may be easier to assess consistently than a patient who is more impaired and falls in the middle or the bottom of a scale for more items across a spectrum of difficulty. It will also be important to examine the role of provider type for this issue: as demonstrated in Section 3.3, there was some statistical evidence that features such as case mix, resources, and procedures have a minor influence on patterns of assessment.

Summary of Findings

- For function items, agreement tended to be higher when evaluating high-functioning patients than when evaluating low-functioning patients.
- The trend toward higher agreement for high-functioning patients held across disciplines, provider types, and the vast majority of function-related items.

Detailed Findings

Fricke and colleagues (1993), in their examination of the FIM®, stratified their analysis of clinician agreement on the functional severity of the patients being rated. RTI adapted this analysis to examine agreement as a function of care provider type and of severity of patient condition.

Table 3-5a presents the percent agreement for each of the function items for which data were collected for both low-functioning and high-functioning patients. Agreement differed significantly between low-functioning and high-functioning patients for each of the items measured (after setting the type I two-tailed error rate to a conservative 0.0025 to adjust for multiple comparisons). For 14 of the 17 items examined here, rates of agreement were higher for high-functioning patients than for low-functioning patients.

Table 3-5a
Proportion of agreement for selected functional items by dependency level of patients

Item	Agreement on low-function patients *n* (%)	Agreement on high-function patients *n* (%)	*t*
VI.A1. Eating	32 (42.7%)	130 (90.9%)	778.51*
VI.A3. Oral hygiene	102 (54.0%)	95 (83.3%)	583.37*
VI.A4. Toilet hygiene	35 (46.7%)	164 (88.2%)	666.30*
VI.A5. Upper body dressing	74 (62.7%)	169 (90.9%)	571.26*
VI.A6. Lower body dressing	105 (89.0%)	133 (71.5%)	−398.04*
VI.B1. Lying to sitting on side of bed	163 (86.2%)	161 (86.6%)	9.04*
VI.B2. Sit to stand	40 (53.3%)	174 (93.6%)	666.37*
VI.B3. Chair/bed-to-chair transfer	135 (71.4%)	142 (76.3%)	108.42*
VI.B4. Toilet transfer	65 (86.7%)	162 (87.1%)	9.29*
VI.B5. Wheelchair use	73 (97.3%)	175 (94.1%)	−127.53*
VI.B5a4. Walk in room once standing	51 (68%)	51 (71.8%)	50.50*
VI.C1. Wash upper body	126 (66.7%)	56 (78.9%)	205.55*
VI.C3. Roll left and right	67 (56.8%)	33 (76.74%)	252.87*
VI.C4. Sit to lying	21 (48.8%)	33 (76.74%)	279.53*
VI.C6. Putting on/taking off footwear	145 (76.7%)	163 (87.6%)	279.12*
VI.C10. Medication management – oral medications	162 (85.7%)	177 (95.2%)	315.76*
VI.C11. Medication management – inhalant/mist medications	111 (94.1%)	135 (72.6%)	−547.13*
Total	1,784 (72.9%)	3,152 (82.2%)	1,095.66*

*$p < .001$.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB07).

Table 3-5b shows overall rates of agreement stratified by provider type, pooling results across all of the function items on CARE. The data indicate that this trend of higher agreement for higher-functioning patients was consistent across all provider types. It is worth noting that percent agreement was similar across provider types when looking at rates of agreement within the two separate patient groups analyzed.

Thus, it appears that the CARE Item Set was more reliable, in terms of interrater agreement, when evaluating high-functioning patients than when evaluating low-functioning patients, and agreement levels appear to be consistent across provider type. However, this result, with respect to higher consistency for higher functioning patients, echoes the conclusions of Fricke and colleagues (1993) in their analysis of the FIM®. It is likely more difficult to find consensus in the evaluation of functionally dependent individuals regardless of the instrument being used in the evaluation.

Table 3-5b
Proportion of agreement for functional items by provider type and level of dependency of patient

Provider type	Agreement on low-function patients *n* (%)	Agreement on high-function patients *n* (%)
Acute	57 (73.1%)	101 (84.2%)
HHA	344 (67.2%)	645 (81.5%)
IRF	786 (78.4%)	1,416 (84.5%)
LTCH	361 (68.4%)	637 (81.7%)
SNF	236 (72.0%)	353 (74.2%)
Total	1,784 (72.9%)	3,152 (82.2%)

NOTE: Acute = acute care facility, HHA = home health agency, IRF = inpatient rehabilitation facility, LTCH = long-term care hospital, SNF = skilled nursing facility.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB07).

This analysis was expanded to examine potential differences between clinicians in different disciplines. The pattern of higher rates of agreement on high-functioning patients also holds across all kinds of clinicians. **Table 3-5c** shows levels of agreement for low- and for high-functioning patients, stratified by clinician discipline. Rates of agreement with the clinical mode were highest, in this analysis, for speech therapists observing high-functioning patients (86.8 percent). Rates of agreement were lowest for each level of dependency for respondents in the "other" category. There was no evidence that agreement rates depend on discipline ($\chi^2(5)$ = 10.7, *n.s.*). Proportions of agreement were even more similar for all named disciplines. Clinicians of varied experience tended to agree with clinical benchmarks more frequently when observing high-functioning patients.

Table 3-5c
Proportion of agreement for functional items by discipline and level of dependency of patient

Discipline	Agreement on low-function patients *n* (%)	Agreement on high-function patients *n* (%)
Case mgr	104 (79.4%)	184 (81.4%)
OT	277 (79.6%)	449 (82.3%)
PT	407 (78.4%)	712 (84.6%)
RN	741 (68.0%)	1,428 (82.7%)
ST	79 (79%)	132 (86.8%)
Other	176 (67.2%)	247 (71.4%)
Total	1,784 (72.9%)	3,152 (82.2%)

NOTE: Case mgr = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, ST = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB08).

To more closely examine the potential interactions between clinician discipline and level of patient independence in function, RTI conducted analyses focusing on six function items representing activities across a spectrum of difficulty for patients to perform, from easy to hard—three of them representing self-care function items, and three representing mobility items (Gage et al., 2012). Generally, agreement was higher for patients who are more independent. **Table 3-5d** and **Table 3-5e** show the agreement rates on each of these items, stratified by clinician discipline and patient functional level. For clinicians of all disciplines, consensus with the clinical mode was less frequent when evaluating low-functioning patients on the "eating" item (VI.A1). Clinicians tended to be most consistent in their ratings of low-functioning, or most dependent, patients with regard to the "putting on/taking off footwear" item (VI.C6). In contrast, ratings were less consistent for the more independent patients for this item. This may be because the more independent patients did not fall as uniformly into the most independent rating, as they likely did for the easier to perform items. As stated previously, this finding, however seems to be consistent with analyses of other assessment instruments (Fricke et al., 1993).

Table 3-5d
Proportion of agreement for selected functional items by discipline and level of dependency of patient for selected self-care function items

Discipline	Eating: Low-function patients *n* (%)	Eating: High-function patients *n* (%)	Upper body dressing: Low-function patients *n* (%)	Upper body dressing: High-function patients *n* (%)	Footwear: Low-function patients *n* (%)	Footwear: High-function patients *n* (%)
Case mgr	0	8 (100%)	4 (66.7%)	11 (100%)	6 (100%)	9 (81.8%)
OT	4 (40%)	18 (94.7%)	9 (52.9%)	24 (92.3%)	17 (100%)	31 (75.6%)
PT	11 (68.9%)	30 (93.8%)	14 (56%)	41 (100%)	25 (100%)	31 (75.6%)
RN	12 (41.4%)	53 (85.5%)	36 (70.6%)	77 (91.7%)	47 (92.2%)	60 (71.4%)
ST	0	8 (100%)	3 (75%)	8 (100%)	3 (75%)	7 (87.5%)
Other	5 (38.5%)	13 (92.9%)	8 (53.3%)	14 (87.5%)	13 (86.7%)	5 (31.3%)
Total	32 (42.7%)	130 (90.9%)	74 (62.7%)	175 (94.1%)	111 (94.1%)	135 (72.6%)

NOTE: Case mgr = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, ST = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB08).

Table 3-5e
Proportion of agreement for selected functional items by discipline and level of dependency of patient for selected mobility function items

Discipline	Roll left and right: Low-function patients *n* (%)	Roll left and right: High-function patients *n* (%)	Sit to lying: Low-function patients *n* (%)	Sit to lying: High-function patients *n* (%)	Toilet transfer: Low-function patients *n* (%)	Toilet transfer: High-function patients *n* (%)
Case mgr	10 (90.9%)	9 (81.8%)	10 (90.9%)	11 (100%)	1 (33.3%)	9 (81.8%)
OT	24 (88.9%)	37 (90.2%)	26 (96.3%)	22 (84.6%)	10 (100%)	15 (93.8%)
PT	30 (75%)	37 (90.2%)	35 (87.5%)	40 (97.6%)	15 (93.8%)	36 (87.8%)
RN	61 (69.3%)	77 (91.7%)	70 (79.6%)	82 (97.6%)	26 (89.7%)	74 (88.1%)
ST	6 (75%)	7 (87.5%)	7 (87.5%)	8 (100%)	4 (100%)	8 (100%)
Other	14 (93.3%)	13 (81.3%)	14 (93.3%)	14 (87.5%)	9 (69.2%)	15 (93.8%)
Total	145 (76.7%)	163 (87.6%)	162 (85.7%)	177 (95.2%)	65 (86.7%)	162 (87.1%)

NOTE: Case mgr = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, ST = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB08).

3.6 Impact of Provider Type, Clinician Discipline, and Patient Functional Status on Direction of Disagreement

Another feature to consider when testing reliability is whether those who disagree with typical or expert assessments tend to do so in any systematic way. A reliable item set should introduce minimal bias in clinician responses; thus, there should be approximately as many item responses that give higher values than the norm as there are responses that give lower values than the norm (and both of these types of responses should be infrequent). To check for systematic biases, RTI used the modal judgments of the clinical expert team as the point of reference and assessed the direction of disagreements with those judgments, that is, whether responding clinicians gave ratings that indicated better or worse health status than did the clinical expert team. This analysis was performed for clinicians of different disciplines and practicing in different provider type settings. It also examined direction of disagreement for different items and for patients of different functional abilities.

Summary of Findings

- Responses tended to indicate better health status for patients than did the assessments of the clinical team.
- The trend of respondents giving more positive assessments held across provider types but was not significant for any individual provider type.
- This trend also held across disciplines but was not significant for speech therapists and OTs.
- This trend does not appear to be affected by how difficult an item is for patients to perform.

An analysis of the direction of discrepancy in judgments by provider type was performed to see if responding clinicians (and subgroups of clinicians) tended to rate certain items more indicative of better or poorer health status than did the group of clinical experts. Such tendencies, if evident, could inform future CARE Item Set training. CARE items were included in this analysis if they met two criteria: (1) they used ordinal-scale values (e.g., the 6-point rating scale for functional status items) and (2) the scale took on a range of three or more possible values (see Appendix A for a list of the variables that were included). For each provider type, and for all provider types combined, RTI used binomial tests, working with the null hypothesis that responses in disagreement with the clinical expert team should be distributed equally in each direction. RTI used the normal approximation to the binomial distribution and derived *z*-scores to describe the direction of discrepancy. A negative *z*-score indicates that ratings tended to be more negative (in health, function, mood, prognosis, etc.) than those of the clinical expert group, whereas a positive score indicates that the ratings tended to indicate better health status than those of the clinical expert group.

Detailed Findings

Table 3-6a shows the results of this analysis. In general, responding clinicians tended to rate the patients in the video more positively than did the clinical expert panel. However, when looking at care provider types separately, the same trend is present but is no longer significant after adjusting the type I error rate for multiple comparisons.

Table 3-6a
Direction of discrepancy by provider type

Provider type	Negative discrepancy† *n* (%)	Agreement	Positive discrepancy‡ *n* (%)	*z*-score
Acute	19 (3.3%)	530 (92.5%)	24 (4.2%)	0.24
HHA	209 (5.2%)	3,496 (86.3%)	348 (8.6%)	1.86
IRF	429 (5.3%)	7,068 (88.0%)	532 (6.6%)	1.05
LTCH	201 (4.9%)	3,483 (85.4%)	393 (9.6%)	2.49
SNF	134 (5.7%)	2,021 (85.8%)	201 (8.5%)	1.16
Total	947 (5.0%)	16,598 (87.0%)	1,543 (8.1%)	3.78*

**p* < .01.

†Response indicates greater functional dependence, more depressed mood, poorer skin integrity, etc., than modal response of clinical experts.

‡Response indicates greater functional independence, less depressed mood, better skin integrity, etc., than modal response of clinical experts.

NOTE: Acute = acute care facility, HHA = home health agency, IRF = inpatient rehabilitation facility, LTCH = long-term care hospital, SNF = skilled nursing facility.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB05).

As with the patterns of agreement discussed previously, the direction of discrepancy tends to be as consistent across respondent discipline as it is between facilities. **Table 3-6b** shows the direction of disagreement between respondents of various disciplines and the modal responses of the RTI clinical expert team. Respondents of each discipline gave more responses that indicated more positive outcomes than did the modal judgments of the clinical team: a binomial test using a normal approximation showed that this difference was significant (after controlling the experiment-wise error rate for multiple comparisons) for case managers, PTs, RNs, and the "other" group. The judgments of speech therapists were most likely to concur with the clinical expert mode with an agreement rate of 89.8 percent and tended to be balanced between more positive and more negative ratings when they disagreed (5.6 percent vs. 4.6 percent). Overall, however, agreement rates were quite high across all disciplines and similar proportions for this subset of items, although the "other" category, as before, had lower rates of agreement.

Table 3-6b
Direction of discrepancy by discipline

Discipline	Negative discrepancy† *n* (%)	Agreement	Positive discrepancy‡ *n* (%)	*z*-score
Case mgr	62 (5.5%)	960 (85.2%)	103 (9.2%)	3.19*
OT	137(5.3%)	2,250 (87.8%)	175 (6.8%)	2.15
PT	148 (3.7%)	3,512 (89.0%)	288 (7.3%)	6.70*
RN	459 (5.2%)	7,700 (86.4%)	749 (8.4%)	8.34*
ST	38 (4.6%)	739 (89.8%)	46 (5.6%)	0.87
Other	103 (6.0%)	1,437 (83.4%)	182 (10.6%)	4.68*
Total	947 (5.0%)	16,598 (87.0%)	1,543 (8.1%)	11.94*

*$p < .01$.

†Response indicates greater functional dependence, more depressed mood, poorer skin integrity, etc., than modal response of clinical experts.

‡Response indicates greater functional independence, less depressed mood, better skin integrity, etc., than modal response of clinical experts.

NOTE: Case mgr = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, ST = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB09).

To take a closer look at the direction of discrepancy as a function of clinician discipline, six functional items were selected for specific analysis (see **Tables 3-6c** and **3-6d**). As described in Section 3.4, these items were selected to represent a set of activities across a range of difficulty for patients to perform (Gage et al., 2012). Three were self-care items (in order of increasing level of difficulty: VI.A1, Eating; VI.A5, Upper body dressing; and VI.C6, Putting on/taking off footwear) and three were mobility items (in order of increasing level of difficulty: VI.C3, Roll left and right; VI.C4, Sit to lying; and VI.B4, Toilet transfer).

Responses to five of these items reflect the tendency of respondents to respond positively to items relative to the modal responses of the clinical expert team. The opposite was true for the easiest mobility item, "Roll left and right." Within items, the general direction of discrepancy does not differ between groups, that is, on the five items where the direction is more positive than negative, the judgments of each group was consistently more positive than negative (and vice versa for "Roll left and right"). The tendency for disagreements in one direction was more

Table 3-6c
Direction of discrepancy for selected functional items by discipline for selected self-care function items

Discipline	Eating: Negative discrepancy† *n* (%)	Eating: Agreement	Eating: Positive discrepancy‡ *n* (%)	Upper body dressing: Negative discrepancy† *n* (%)	Upper body dressing: Agreement	Upper body dressing: Positive discrepancy‡ *n* (%)	Footwear: Negative discrepancy† *n* (%)	Footwear: Agreement	Footwear: Positive discrepancy‡ *n* (%)
CM	0	17 (77.3%)	5 (22.7%)	0	21 (75%)	7 (25%)	3 (10.7%)	20 (71.4%)	5 (17.9%)
OT	1 (2.0%)	38 (77.6%)	10 (20.4%)	7 (10.9%)	50 (78.1%)	7 (10.9%)	5 (7.8%)	53 (82.8%)	6 (9.4%)
PT	6 (7.4%)	67 (82.7%)	8 (9.9%)	1 (1.0%)	75 (75.8%)	23 (23.2%)	5 (5.1%)	76 (76.8%)	18 (18.2%)
RN	8 (4.7%)	128 (74.4%)	36 (20.9%)	3 (1.4%)	160 (72.7%)	57 (25.9%)	12 (5.5%)	182 (83.9%)	23 (10.6%)
ST	1 (5.3%)	14 (73.7%)	4 (21.1%)	1 (5%)	15 (75%)	4 (20%)	0	18 (90%)	2 (10%)
Other	1 (2.9%)	24 (68.6%)	10 (28.6%)	3 (7.0%)	33 (76.7%)	7 (16.3%)	1 (2.3%)	28 (65.1%)	14 (32.6%)
Total	17 (4.5%)	288 (76.2%)	73 (19.3%)	15 (3.2%)	354 (74.7%)	105 (22.2%)	26 (5.5%)	377 (80.0%)	68 (14.4%)

†Response indicates greater functional dependence, more depressed mood, poorer skin integrity, etc., than modal response of clinical experts.

‡Response indicates greater functional independence, less depressed mood, better skin integrity, etc., than modal response of clinical experts.

NOTE: CM = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, ST = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB09).

Table 3-6d
Direction of discrepancy for selected functional items by discipline for selected mobility function items

Discipline	Roll left and right: Negative discrepancy† *n* (%)	Roll left and right: Agreement	Roll left and right: Positive discrepancy‡ *n* (%)	Sit to lying: Negative discrepancy† *n* (%)	Sit to lying: Agreement	Sit to lying: Positive discrepancy‡ *n* (%)	Toilet transfer: Negative discrepancy† *n* (%)	Toilet transfer: Agreement	Toilet transfer: Positive discrepancy‡ *n* (%)
CM	5 (15.2%)	25 (75.8%)	3 (9.1%)	0	27 (96.4%)	1 (3.6%)	2 (8%)	20 (80%)	3 (12%)
OT	11 (15.1%)	59 (80.8%)	3 (4.1%)	4 (6.3%)	59 (92.2%)	1 (1.6%)	0	53 (93.0%)	4 (7.0%)
PT	9 (7.0%)	104 (91.2%)	2 (1.8%)	0	93 (93.9%)	6 (6.1%)	0	84 (93.3%)	6 (6.7%)
RN	27 (10.5%)	213 (82.9%)	17 (6.6%)	1 (0.5%)	200 (93.02%)	14 (6.5%)	1 (0.5%)	179 (90.9%)	17 (8.6%)
ST	3 (12.5%)	21 (87.5%)	0	0	19 (100%)	0	1 (5%)	19 (95%)	0
Other	1 (2.3%)	38 (88.4%)	4 (9.3%)	1 (2.3%)	40 (93.0%)	2 (4.7%)	1 (2.4%)	36 (87.8%)	4 (9.8%)
Total	55 (10.1%)	460 (84.6%)	29 (5.3%)	6 (1.3%)	438 (95.6%)	24 (5.1%)	5 (1.2%)	391 (90.9%)	34 (7.9%)

†Response indicates greater functional dependence, more depressed mood, poorer skin integrity, etc., than modal response of clinical experts.

‡Response indicates greater functional independence, less depressed mood, better skin integrity, etc., than modal response of clinical experts.

NOTE: CM = case manager, OT = occupational therapist, PT = physical therapist, RN = registered nurse, ST = speech therapist, and Other includes licensed practical nurses.

SOURCE: RTI analysis of CARE Item Set from video reliability testing, 2009 (RTI program reference: DB09).

pronounced for some items. For example, relatively high proportions of respondents indicated more independence on the “Eating” and “Upper body dressing” items. However, there appears to be no evidence that either item difficulty or clinician discipline have any systematic effect on the direction of disagreement.

[This page intentionally left blank.]

SECTION 4
CONCLUSION

Medicare beneficiaries receive post-acute care from a variety of provider types, each featuring a different but overlapping range of services. While 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 CARE Item Set is a standardized set of items designed to address this issue by providing a set of items that could be used in acute and post-acute settings. It draws on the science behind several currently mandated instruments, such as IRF-PAI, MDS, and OASIS, as well as input from stakeholders, technical experts, and public comment. To compare health outcomes across provider types, a standardized assessment must not only comprise items that are relevant to all care provider types but also must elicit consistent responses from clinician assessors in different provider types. Video reliability tests were designed to allow respondents of various disciplines and experience and from different provider types to evaluate the same set of patients using the CARE Item Set. These responses were analyzed to measure the level of clinician agreement across provider type settings and to identify any patterns of agreement that might inform the future development of the tool.

Video clips (each approximately 20 minutes in length) depicting nine "patients" were created by RIC. These patients were evaluated using the CARE Item Set by clinical experts from RTI and VNS-NY. These video clips were evaluated by clinicians associated with 28 providers enrolled in PAC-PRD data collection. RNs represented the largest subgroup of respondents (47 percent), followed by PTs (21 percent), OTs (14 percent), and the "other" group, which was largely composed of LPNs (8 percent). Among the different provider types, 43 percent of assessments were contributed by IRFs, 22 percent by HHAs, 21 percent by LTCHs, and 12 percent by SNFs, and 3 percent came from acute hospitals. Video reliability analyses assessed rates of agreement between groups of respondents and the modal responses of the clinical expert team.

Rates of agreement on all items were similar among RNs, PTs, and OTs, regardless of whether the clinicians were familiar with given CARE items through their customary assessment practices during the PAC-PRD data collection. However, the "other" group was significantly different from all other disciplinary groups. It appears that the majority of clinicians assessed the patients in ways that were consistent with each other and with the team of clinical experts. Further, there was no evidence to suggest that clinicians who did not typically assess various items had any particular issues with providing consistent ratings on those items. Thus, it appears from these analyses that clinicians can be trained to reliably use the CARE Item Set.

Clinicians practicing in different provider settings had similar rates of agreement across provider type, except for clinicians practicing in SNFs, who tended to have higher rates of agreement with the RTI clinical expert team than providers practicing in other provider type settings. It is unclear why the clinicians practicing in SNFs had higher rates of agreement than those practicing in other settings. Generally, however, rates of agreement were high across provider types, indicating good reliability across settings.

For function items in the CARE Item Set (Section VI), rates of agreement tended to be higher when clinicians assessed high-functioning patients relative to when they assessed low-functioning patients. This trend held across most items in the function section and across all provider types. Note that a similar trend was found by Fricke and colleagues (1993) in their examination of similar measures of function used in a different assessment. Thus, there might be difficulty achieving consensus among clinicians inherent in evaluation of low-functioning patients, regardless of the assessment being used.

When responding clinicians disagreed with the clinical expert team, they tended to evaluate items in ways that indicated more positive health status (e.g., rating mood as better or functioning level as better). However, this difference with the clinical expert team was significant only when assessments were examined overall. No provider type showed a significant difference by itself.

In general, it appears that the CARE Item Set is reliable when used across different provider types, when used by different types of clinicians, and regardless of whether clinicians typically fill out particular items in their ordinary assessment practice. While the set of video analyses performed here generally supports the reliability of the CARE Item Set, the analyses raised some minor concerns, including the higher rates of agreement among clinicians practicing in SNFs than those practicing in other PAC settings; the difficulty in getting consistent ratings for low-functioning patients, though this is not an issue unique to CARE; and the tendency of respondents to overestimate the quality of patient health relative to clinical experts, though this pattern did not vary by provider type.

REFERENCES

Fricke, J., Unsworth, C., and Worrell, D.: Reliability of the functional independence measure with occupational therapists. Aust. Occup. Ther. J. 40(1):7-15, 1993.

Gage, B. J., Morley, M., Constantine, R., et al.: Examining Relationships in an Integrated Hospital System. Mar. 2008. Retrieved August 14, 2012, from http://aspe.hhs.gov/health/reports/08/examine/report.html.

Gage, B., Smith, L., Ross, J., et al.: Volume 2 of The Development and Testing of the Continuity Assessment Record and Evaluation (CARE) Item Set: Final Report on Reliability Testing. Waltham, Mass. RTI International, August 2012.

[This page intentionally left blank.]

APPENDIX A: CARE ITEM SET VARIABLES INCLUDED IN DIRECTION OF DISAGREEMENT ANALYSES

The following variables were included in the analysis profiled in Section 3.6, which considered whether clinicians disagreed in systematic ways dependent on their discipline or the provider type where they practice:

II.B5a	Self-care	V.G1a	Mobility endurance
II.B5b	Mobility (ambulation)	V.G1b	Sitting endurance
II.B5c	Stairs (ambulation)	VI.A1	Eating
II.B5d	Mobility (wheelchair)	VI.A2	Tube feeding
II.B5e	Functional cognition	VI.A3	Oral hygiene
III.G1	Presence of pressure ulcers	VI.A4	Toilet hygiene
III.G2a	Number of stage 2 pressure ulcers	VI.A5	Upper body dressing
III.G2b	Number of stage 3 pressure ulcers	VI.A6	Lower body dressing
III.G2c	Number of stage 4 pressure ulcers	VI.B1	Lying to sitting
III.G2d	Number of unstageable pressure ulcers	VI.B2	Sit to stand
III.G2e	Number of unhealed stage 2 ulcers known to be present for more than 1 month	VI.B3	Chair/Bed-to-chair transfer
IV.B3a	Repetition of three words	VI.B4	Toilet transfer
IV.B3b1	Identification of month	VI.B5a1	Walk 150 feet
IV.B3b2	Identification of day	VI.B5a2	Walk 100 feet
IV.B3c1	Patient recalls “sock”	VI.B5a3	Walk 50 feet
IV.B3c2	Patient recalls “blue”	VI.B5a4	Walk once standing
IV.B3c3	Patient recalls “bed”	VI.B5b1	Wheel 150 feet
IV.D1	CAM: Inattention	VI.B5b2	Wheel 100 feet
IV.D2	CAM: Disorganized thinking	VI.B5b3	Wheel 50 feet
IV.D3	CAM: Altered level of consciousness/alertness	VI.B5b4	Wheel once seated
IV.D4	CAM: Psychomotor retardation	VI.C1	Wash upper body
IV.F2b	Little interest or pleasure in doing things: number of days	VI.C2	Shower/bathe self
IV.F2d	Feeling down, depressed, or hopeless: number of days	VI.C3	Roll left and right
IV.F3	Feeling sad: Frequency	VI.C4	Sit to lying
V.A3a	Frequency of bladder incontinence	VI.C5	Picking up object
V.A3b	Frequency of bowel incontinence	VI.C6	Putting on/taking off footwear
V.C1a	Understanding verbal content	VI.C7a	One step (curb)
V.C1b	Expression of ideas and wants	VI.C7b	Walk 50 feet with two turns
V.C1c	Ability to see in adequate light	VI.C7c	Twelve steps–interior
V.C1d	Ability to hear	VI.C7d	Four steps–exterior
V.E1a	Grip strength: Left hand	VI.C7e	Walking 10 feet on uneven surfaces
V.E1b	Grip strength: Right hand	VI.C7f	Car transfer
V.F1a	Respiratory status with supplemental oxygen	VI.C7h	Wheel long ramp
V.F1b	Respiratory status without supplemental oxygen		