

4.2.3a Additional Reliability Testing Results: Data Element Reliability Testing

Overview of Data Element Reliability Testing

The goal of reliability testing is to ensure that data elements (i.e., items) obtain consistent results when administered or used by different clinicians. The functional status items initially underwent reliability testing at the item- and scale-level in multiple types of providers in conjunction with the Post-Acute Care Payment Reform Demonstration. Item-level testing included inter-rater reliability testing within provider and the use of videotaped standardized patients for inter-rater reliability testing across provider/care settings. A brief summary of this testing is provided below. The full reports describing the testing are available at <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Post-Acute-Care-Quality-Initiatives/CARE-Item-Set-and-B-CARE.html>. The main reports are:

1. Smith LM, Deutsch A, Hand LB, Etlinger AL, Ross J, Abbate JH, Gage-Croll Z, Barch D, Gage BJ. (September, 2012). *Continuity Assessment Record and Evaluation (CARE) Item Set: Additional Provider-Type Specific Interrater Reliability Analyses*. Prepared for Centers for Medicare & Medicaid Services.
2. Smith LM, Deutsch A, Barch D, Ross J, Shamsuddin KM, Abbate JH, Schwartz C, Gage BJ. (September, 2012). *Continuity Assessment Record and Evaluation (CARE) Item Set: Video Reliability Testing*. Prepared for Centers for Medicare & Medicaid Services.

Traditional Inter-rater Reliability Study

The reliability of the functional items was tested in a subset of 34 providers, including acute care hospitals, HHAs, IRFs, LTCHs, and SNFs, distributed across 11 geographic areas. Each provider completed a duplicate CARE Item Set (admission or discharge assessment) on 10–20 patients included in the Post-Acute Care Payment Reform Demonstration, in accordance with the guidelines and protocols.

Providers were asked to enroll a convenience sample of a set number of Medicare patients each month, representing a range of function and acuity. The overall patient sample size for each of the functional items was 450 for self-care items and 449 for mobility items (448 for transfers). After exclusions for missing data (unknown/not attempted/inapplicable), the effective sample sizes for the reliability testing were as follows:

- Eating: 401
- Oral hygiene: 414
- Toileting hygiene: 416

- Lying to sitting on the side of the bed: 412
- Sit to stand: 387
- Chair/bed to chair transfer: 392
- Toilet transfer: 361
- Walk 10 feet: 52
- Wheel 10 feet: 46

The inter-rater reliability study included patients who were assessed by two different clinicians (raters), and the agreement of the clinicians' rating was calculated. Clinicians were instructed to have pairs of raters complete both patient assessments at the same time. Responses to items were obtained by direct observation of the patient by the clinician, and occasionally, supplemented by one or more of the following predetermined, matched methods: patient interviews (with each team member taking turns conducting and observing patient interviews); interviews with relatives/caregivers of the patient for certain items; and/or 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 function 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 items that required hands-on assistance. Raters were instructed not to discuss item scoring during the assessment, nor to share item scores until the data were entered into the study database and finalized. Providers submitted data via the online CARE application for both assessments in each pair.

For categorical items, kappa statistics (kappa) indicate the level of agreement between raters using ordinal data, taking into account the role of chance agreement. The ranges commonly used to judge reliability based on kappa are 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 0.81–1.00 = almost perfect.

For categorical items with only two responses available, researchers calculated only unweighted kappas. For items with more than two responses, researchers calculated both weighted and unweighted kappas. Unweighted kappa assumes the same "distance" between every one-unit difference in response across an ordinal scale. Researchers 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.^{1,2} Fleiss-Cohen weights put lower emphasis on disagreements between responses that fall near 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 play a role in the results. Hence, we report both weighted and unweighted kappas to give the range of agreement found under the two sets of assumptions.

Additionally, researchers 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 the traditional reliability study, kappa statistics indicated substantial agreement among raters. The weighted kappa values for the self-care items range between 0.798 for eating to 0.869 for upper-body dressing. Unweighted kappas ranged from 0.598 for oral hygiene to 0.634 for upper-body dressing. Provider-specific analyses of core self-care items show similar agreement to the overall estimates. The lower-body dressing item had the highest overall weighted kappa (0.855), whereas the eating item had the lowest (0.798). Unweighted overall kappas ranged from 0.636 (toileting) to 0.598 (oral hygiene). Acute hospitals had the highest weighted kappas across all self-care items.

The weighted kappa values for the mobility items ranged between 0.878 for toilet transfers to 0.901 for sit to stand and chair/bed to chair transfer. Unweighted kappas ranged from 0.667 for walk 10 feet to 0.762 for sit to stand. Provider-specific analyses of core mobility items show similar agreement to the overall estimates. The sit-to-stand and chair transfer items both had a weighted kappa of 0.901, whereas the lying to sitting item had a weighted kappa of 0.855. Unweighted overall kappas ranged from 0.693 (lying to sitting) to 0.762 (sit to stand).

Videotaped Standardized Patients Reliability Study

For the video reliability study, which was designed to examine the level of clinician agreement across care settings, clinicians in each setting were asked to assess “standardized” patients presented through a videotape of a patient assessment. This ensured that the same information was presented to each clinician and allowed examination of differences in scoring effects among different clinicians examining the “same” patient.

The patient “case studies” in each of the videos varied in terms of medical complexity, functional abilities, and cognitive impairments. The nine videos included patients classified as high, medium, or low ability/complexity for each of these three areas. Each facility or agency received three videos, one of which demonstrated one of the following elements: cognitive impairments, skin integrity problems, a wheelchair-dependent patient, and a variety of mid-level functional activities. The mid-level functional activities were considered to be the most challenging for clinicians to score and are thus of particular interest in establishing reliability. Each clinician involved in the video study watched three videos and assessed the patients according to the study guidelines and protocols. Each video was approximately 20 minutes long and had a corresponding item set arranged in the sequence in which the items appeared in the video.

The sample included 28 providers (550 assessments), which included 3 acute hospitals (15 assessments [3%]); 9 HHAs (118 assessments [22%]); 8 IRFs (237 assessments [43%]); 3 LTCHs (114 assessments [21%]); and 5 SNFs (66 assessments [12%]). Participating providers included case managers (6% of assessments), occupational therapists (14% of assessments), physical therapists (21% of assessments), registered nurses (47% of assessments), speech therapists (5% of assessments), and others, mostly licensed practical nurses (LPNs; 8% of assessments).

Two main analytic approaches were used for assessing the video reliability of the CARE items, adhering closely to the methods used by Fricke et al.³ in their video reliability study of a functional assessment instrument. First, percent agreement with the mode response was calculated for each CARE item included in at least one of the nine videos. Unlike the approach used by Fricke et al., researchers did not consider agreement at one response level above and below the mode, and instead used a stricter approach looking at direct modal agreement only. In the second approach, percent agreement with the internal clinical team's consensus response was also calculated. This second measure not only gives an indication of item reliability, but also reflects training consistency for the providers.

The video reliability study indicated substantial agreement with the mode and clinical team among all items, typically upwards of 70%. The notable exception to this trend exists among the clinicians in the "Other" category (mostly LPNs); they consistently had the lowest levels of agreement among all core self-care items, ranging from 50 to 72%. For the toileting hygiene item, the agreement with the clinical team was lower than with the mode. This occurred because the clinical team response differed from the mode for these three items in either one or two videos. Nonetheless, because the clinical team response and mode were identical on most of the videos, agreement was still quite high for these items. In general, study clinicians had responses on average that agreed with the expert clinical team or were slightly lower.

The video reliability study indicated substantial agreement with the mode and clinical team for the lying-to-sitting, sit-to-stand, chair/bed to chair transfer, and toilet transfer items (greater than 76%). Although rates of agreement with the mode and clinical team response were generally identical, for the toilet transfer item, the clinical team agreement is slightly lower. The items for walking and wheeling distances showed more variable levels of agreement across disciplines, with overall agreement generally in the moderate range (50–78%). For the Walk 10 feet item, there was a notable decrease in the agreement with the clinical team compared to agreement with the mode. This occurred because in two of the four videos where this item was assessed, the clinical team response differed from the mode.