Original Study

Making the Investment Count: Revision of the Minimum Data Set for Nursing Homes, MDS 3.0

Debra Saliba MD, MPH a,*, Joan Buchanan PhD b

a UCLA/Jewish Home Borun Center for Gerontological Research, Los Angeles, CA; Greater Los Angeles VA GRECC and HSR&D Center of Excellence; RAND, Santa Monica, CA
b Department of Health Care Policy (retired), Harvard Medical School, Boston, MA

Keywords: Nursing home, Minimum Data Set assessment, reliability, efficiency, preferences, consumers, response rate, interviews

ABSTRACT

Background: The Minimum Data Set (MDS) is a potentially powerful tool for implementing standardized assessment in nursing homes (NHs). Its content has implications for residents, families, providers, researchers, and policymakers, all of whom have expressed concerns about the reliability, validity, and relevance of MDS 2.0. Some argue that because MDS 2.0 fails to include items that rely on direct resident interview, it fails to obtain critical information and effectively disenfranchises many residents from the assessment process.

Purpose: Design a major revision of the MDS, MDS 3.0, and evaluate whether the revision improves reliability, validity, resident input, clinical utility, and decreases collection burden.

Design and Methods: In the form design phase, we gathered information from a wide range of experts, synthesized existing literature, worked with a national consortium of VA researchers to revise and test eight sections, pilot tested a draft MDS 3.0 and revised the draft based on results from the pilot. In the national validation and evaluation phase, we tested MDS 3.0 in 71 community NHs and 19 VHA NHs, regionally distributed throughout the United States. The sample was selected based on scheduled MDS 2.0 assessments. Comatose residents were excluded. A total 3822 residents of community NHs in eight states were included. The evaluation was designed to test and analyze inter-rater agreement (reliability) between research nurses and between facility staff and research nurses, validity of key sections, response rates for interview items, anonymous feedback on changes from participating nurses, and time to complete the MDS assessment.

Results: The reliability for research nurse to research nurse and for research nurse to facility staff was good or excellent for most items. Response rates for the resident interview sections were high: 90% for cognitive, 86% for mood, 85% for preferences, and 87% for pain. Staff survey responses showed increased satisfaction with clinical relevance, validity and clarity compared with MDS 2.0. The test version of the MDS 3.0 took 45% less time for facilities to complete.

Implications: Improving the reliability, accuracy, and usefulness of the MDS has profound implications for NH care and public policy. Enhanced accuracy supports the primary legislative intent that MDS be a tool to improve clinical assessment and supports the credibility of programs that rely on MDS.

* Address correspondence to Debra Saliba, MD, MPH, UCLA JH Borun Center for Gerontological Research, 10945 Le Conte Ave, Suite 2339, Los Angeles, CA 90095.
E-mail address: saliba@rand.org (D. Saliba).

In 1986, the Institute of Medicine issued its landmark report on quality of care in United States nursing homes (NHs). It recommended shifting the nation’s strategy for improving NH care from structural evaluations of NHs to systematic, standardized assessments of resident’s cognitive, functional, and emotional needs. Such assessments were seen as a crucial foundation for developing appropriate care plans. The subsequent passage of NH legislation in the 1987 Omnibus Budget Reconciliation Act was a seminal event in NH policy. The legislation mandated the development of a resident assessment instrument describing important domains of resident health and quality of life.
The Resident Assessment Instrument (RAI), centered on the Minimum Data Set (MDS), was introduced into community NHs in 1991. It was later revised (MDS 2.0) in 1998, and in 1998, the Veterans Administration (VA) implemented the MDS. The MDS contains items designed to assess NH residents’ functional status, mood, and medical conditions. MDS data are used to develop quality indicator reports and to determine Medicare post acute care payment. Some states also use the MDS for case-mix based Medicaid reimbursement.

Thus, MDS content has implications for residents, families, providers, and policymakers. These stakeholders have over 18 years of experience with MDS and give it mixed reviews. Its introduction has been temporally associated with improvements in some outcomes and care processes. Some scales that researchers can calculate from MDS data perform well. MDS reliability and validity. One issue has been the difference between the tool’s performance in ideal circumstances and its performance when used by actual NH staff. The MDS has overall acceptable inter-rater reliability when collected by trained research nurses whose only focus is data collection. However, comparisons of trained research nurse ratings to facility-nurse ratings have been mixed. Some studies show acceptable reliability for some items; others show important disagreement, and considerable variation in agreement across items. Many are also concerned that the length and collection burden of MDS exacerbate problems with data quality and validity while redirecting limited resources from important care.

Concerns also have been voiced about how MDS items relate to the physical and emotional domains of health and quality of life. Several important MDS sections fail to show acceptable validity. In addition, critics argue that because the MDS does not include items that require direct questioning of residents, it fails to obtain critical information about resident quality of life.

To some extent, these issues relate to the challenges of designing an instrument to evaluate NH populations who have high disease burden, functional dependence, and risk for serious declines in health. High levels of cognitive impairment (CI) present an additional challenge. To ensure assessment across all cognitive levels, MDS developers designed items that, theoretically, could be completed based solely on staff observation and chart review. Although the instruction manual encouraged obtaining the resident’s input, chart review became the common default option in many NHs. This reliance on chart review may be exacerbated by retrospective “quality” audits that require matching information in medical records. This approach inadvertently resulted in excluding the voice of an estimated 50% or more of NH residents, including some with medical records. This approach inadvertently resulted in excluding the retrospective option in many NHs. This reliance on chart review may be exacerbated by information about resident quality of life.

In response to these concerns, major efforts to revise MDS was initiated. The remainder of this article describes the goals, methods, and overall results of a national research effort to develop version 3.0 of the Minimum Data Set (MDS 3.0).

**Goals for Revising MDS**

Our overall objectives were to provide scientific input to improve the accuracy of MDS assessments and to enhance its clinical usefulness. Early in the project, we worked with the Centers for Medicare & Medicaid Services (CMS) and stakeholders to identify the salient issues for revising the MDS. This process identified five basic goals:

1. Improve MDS clinical relevance and accuracy. This goal supports the primary legislative intent that MDS be a tool to improve clinical assessment and thereby care quality in NHs.
2. Increase the voice of the resident. This goal directly relates to enhancing MDS relevance and moving toward improved assessment and resident-centered care.
3. Improve user satisfaction. This goal recognizes that provider attitudes are key determinants of quality improvement implementation. Negative provider attitudes toward the MDS 2.0 are often cited as a reason that NHs have not fully integrated it into care planning.
4. Increase the efficiency of reports by obtaining useful information with the least possible provider burden.
5. Maintain CMS’s ability to use MDS data for quality measurement and payment (resource utilization groups [RUGs] classification).

**Methods**

Development and testing of MDS 3.0 began in 2003 and concluded in 2008. In the form development phase, we (1) gathered information from the literature and a wide range of stakeholders and other experts, (2) worked with a national consortium of VA researchers to revise and test eight MDS sections, (3) integrated the results of these activities to pilot test a draft MDS 3.0, and (4) revised the draft MDS 3.0 based on pilot results. The national testing phase tested the resulting MDS 3.0 in 71 community and 19 VA NHs.

**Methods to Develop Test Form**

**Stakeholder and Expert Feedback on MDS 2.0 and Proposed MDS 3.0**

This section presents some of our methods for obtaining stakeholder and expert feedback on revising MDS. In addition to those described here, we reviewed the published literature to identify advances in assessment science for older adults, conducted a Town Hall meeting, worked with CMS representatives, contacted content experts for input, and assembled teleconference workgroups on specific topics.

**Matrix of Written Comments**

CMS posted a draft MDS 3.0 for public comment. More than 200 questions and 1000 comments were received from 144 different groups or individuals. We performed a content analysis of these comments at an item level. Approximately a fourth were suggestions for modification of items, another 20% were suggestions for new or additional items, and around 10% were suggestions for items to delete.

**Technical Expert Panel**

The Commonwealth Fund provided a grant to convene a technical expert panel (TEP). Forty-five groups nominated over 150 individuals for this TEP or the subsequent validation panel (described below). From these nominees, we identified panels representing a wide range of experience in NH care delivery, management, and quality improvement across MDS items. Members are listed in the online Appendix 1 (available at www.jamda.com).

In a 2-day meeting, the TEP identified the most important revision goals as improving the function of MDS as a clinical tool and enhancing its efficiency to screen for important issues. The TEP emphasized that it was more important to have items be clear and relevant than to have them be short or on fewer pages. The TEP recommended that the MDS focus on items that would improve initial screening for common and often-missed geriatric syndromes.

The TEP rated the importance of 52 constructs in the MDS 2.0 or draft 3.0 for (a) care of persons requiring basic NH services, (b) care of persons requiring skilled nursing or rehabilitation after acute illness, (c) costs or resource use, and (d) understanding facility quality. Each TEP member ranked the constructs using a 5-point Likert scale from 1 (not at all important) to 5 (extremely important). The overall mean
for the TEP's final ratings of the clinical importance of the constructs did not differ significantly for long-stay vs post-acute residents (4.0 vs 3.8). The sole exception was "estimated length of stay," which was judged important only for post-acute care. For nearly every construct, the TEP’s rating of clinical importance out-ranked importance for cost and quality measurement. The eight clinical constructs ranked most important were: pain assessment, falls, cognitive function, activities of daily living, behavior, delirium, continence, and pressure ulcers.

The four concepts rated least important were: time awake, past roles, number of physician orders, and number of physician visits.

Validation Panel

The Validation Panel (listed in online Appendix 1 at www.jamda.com) used a modified-Delphi process to rate the validity and feasibility of potential items. This expert panel methodology is a well-studied quantitative approach that synthesizes the scientific literature and current expert knowledge in order to specify appropriate measures. It shows acceptable inter-panel reliability and validity. It is particularly useful when research findings must be translated from narrowly focused studies to larger populations.

Specifically, we identified 438 proposed MDS items or scales based on a literature review and contacts with content experts. We provided the panel with a literature synthesis for proposed items, MDS 2.0 item reliability and a summary of written feedback and TEP input. The panel was asked to rate validity based on whether an item or scale was accurate, sensitive for identifying the target condition, specific and explicit, and important as a care planning link. An item was feasible to collect based on staff’s ability to accurately complete the item, the reasonableness of training requirements, and whether the needed staffing resources were consistent with those found in an average community NH.

The validation panel voted on the items by confidential ballot. Panelists could modify or suggest additional items. The initial voting was followed by a 2-day meeting, after which panelists re-voted. We analyzed the panel rating for each item. An item was considered valid if the median validity rating was in the valid range and the panelists’ votes showed statistically significant agreement. If the median rating for an item was in the valid range but a significant number of panelists voted the item not valid, then disagreement was noted and the item was not considered valid. The feasibility votes were treated likewise. We used the results to select items for VA testing (below) and for the pilot MDS.

Veterans Administration Design of Pilot and Validation Activities

As a result of stakeholder and expert input, we identified eight key sections of MDS needing extensive additional revision and testing: mental status, diagnostic coding, delirium, pain, falls, depression, behavior disorders, and quality of life. In October, 2004, the VA Health Services Research and Development Service (VA HSR&D) funded a research proposal that aimed to contribute to revising these sections. This research allowed extensive item development and testing before conducting the national trial in VA and community facilities. During the VA work, the national CMS project continued to revise other items based on expert feedback and literature review.

Through the VA HSR&D project, a research consortium (shown in Appendix 1 at www.jamda.com) worked with the principal investigators of the research and evaluation team to develop, pilot test key MDS sections and related validation protocols. The VA pilot work yielded important findings for several MDS sections.

- Mental status assessment: NH staff could use a simple performance-based screen, simplifying assessment and collecting cognitive items with greater recognition in other settings.
- Delirium: A standardized delirium assessment, the Confusion Assessment Method (CAM), validated in older hospitalized adults is feasible for use in NHs.
- Mood: Direct resident interview for signs and symptoms of depression is feasible, even in residents with moderate cognitive impairment (CI). The VA research showed that NHs could use the PHQ-9, a validated mood assessment used in other settings.
- Behavior: Item modifications met concerns expressed by consumers and providers about the need for language that avoids stigmatization.
- Quality of Life (QoL): The Preference Assessment Tool (PAT) was developed to systematically solicit resident preferences related to QoL domains, including activities. The longer pilot version of the PAT averaged 8 minutes to administer. Residents with moderate CI could provide stable responses to questions about the importance of particular QoL domains and activities.
- Balance: Refined items could guide NH staff in reliably identifying components of gait and transitions that relate to fall risk.
- Diagnoses: Diagnostic categories and diagnoses relevant to NH resident care planning were identified using prevalence data and expert input. Enhanced algorithms for identifying active diagnoses improved agreement between research nurses and clinical nurses.
- Pain: Direct resident interview about pain is feasible, even in residents with moderate to moderately severe CI, a finding consistent with prior NH studies. Serial surveys of residents with varying levels of CI found that residents were able to recall if they had had pain in the preceding 5 days. Resident report of the effect of pain on daily function added information to severity ratings.
- Falls: A revised MDS fall item had improved sensitivity for detecting falls. Facility-nurses were able to use a revised item that asks about fall-related injury to accurately code fall case studies.

MDS 3.0 National Pilot Test

We translated the results from stakeholder and expert input, literature review, workgroups on specific topics and VA pilot testing into a revised form, developed community-based protocols for validation and reliability testing, and developed draft instructions for new MDS items. A Workgroup on the Integrated Tool (WIT) reviewed these items. Participants are listed in the online Appendix 1 (available at www.jamda.com). WIT feedback was incorporated into additional MDS 3.0 item and form revisions. Items were also added or modified based on the needs of a CMS RUGs recalibration study.

We next pilot tested the draft MDS 3.0 in four VA facilities, five Colorado NHs, and one hospital-based transitional care unit. At the end of pilot testing, the pilot study staff provided written feedback and participated in a conference call to review the feedback. Based on this feedback, we made additional revisions and finalized items, instructions, and validation protocols for testing in a national NH sample.

Final Content of MDS 3.0 for National Testing

The most significant proposed changes were to use direct resident interviews to assess cognition, mood, pain, preferences for customary routines and daily activities, and to ask about interest in speaking with someone about return to community. The form told the assessor to attempt to interview every resident who was capable of making
his/herself understood at least some of the time (MDS 2.0 item C4). For a resident who could not make his/herself understood, or who could not complete an attempted interview, alternative staff observation items were provided. Interviews were to be conducted on the day before, day of or day after the assessment reference day.

Other significant changes were to: regroup and revise the behavior items; redefine dependence and assess most dependent episode for ADLs; change balance items to focus on balance during transitions that present high fall risk; revise fall reporting; revise toileting item to collect information on results of urinary toileting trials; remove catheter from definition of "continent"; report pain management; include signs and symptoms of a swallowing disorder; and revise the pressure ulcer section to reflect NPUAP definitions, eliminate reverse staging, document ulcer size and document presence on admission. Algorithms for identifying active diagnoses were also incorporated.

The draft MDS used a 5-day assessment look-back for most items. The exceptions were mood (2 weeks), diagnoses, mobility prior to admission, trial of urinary toileting program, falls, height and weight, healed pressure ulcers, special treatments, influenza and pneumococcal vaccine, therapies, and nursing rehabilitation/restorative care.

We also modified the form design. The redesign focused on developing consistent cognitive maps and layout for items and responses in order to increase clarity and ease of use. Form redesign included larger fonts, logical page breaks, consistent patterns for response types, fewer items per page, and more instructions and definitions on the form.

Methods for National Test of Draft MDS 3.0

The national validation and evaluation of the MDS 3.0 included 71 community NHs (3822 residents) and 19 VA NHs (764 residents), regionally distributed throughout the United States. The evaluation was designed to support testing of inter-rater reliability between research nurses and between facility staff and research nurses, validity of key sections, time needed to complete the MDS, and anonymous survey feedback from participating assessors. Methods varied between the community and VA samples. For simplicity, this section describes the national MDS 3.0 test in the community sample.

Selection of States and Facilities

We used the network of Quality Improvement Organizations (QIOs) to facilitate recruitment of research nurses and NHs for the national data collection. Eight QIOs participated: California, Colorado, Georgia, Illinois, New Jersey, North Carolina, Pennsylvania, and Texas. The QIOs recruited 16 research nurses (two per state). We asked each QIO to recruit 10 NHs for the state trainings. The goal was to have 70 NHs in the national sample with a back-up NH trained in each state. One NH changed ownership and withdrew midway through the study and was replaced by an alternate; both facilities contributed data to the study. Sixty-three percent of the NHs were for profit facilities, 35% were hospital based and 22% described themselves as rural. The participating NHs ranged in size from less than 50 beds (4%) to 250 or more beds (22%), with reported size of 100–149 as the largest group (35%). Thirty-five percent had a dedicated Alzheimer's unit. Online Appendix 2 (available at www.jamda.com) provides additional detail about the recruited facilities.

Facility-Nurse Data Collectors

Participating NHs identified the person who was primarily responsible for completing the MDS in their facility. This facility-nurse underwent training on MDS 3.0 and led data collection in the facility. Online Appendix 3 (available at www.jamda.com) shows the self-reported characteristics of the nurses based on their responses to a questionnaire at the end of the study.

National MDS 3.0 Training: Train the Trainer

The 16 research nurses attended two 4-day training sessions. The first session provided training on the items, and obtained feedback on item format and instruction clarity. During the training, the research nurses collected the MDS 3.0 items on two to three residents in local NHs. We revised the item format and instructions based on the research nurses' feedback.

The second session reviewed revisions to instructions and trained research nurses on the validation protocols. Mental health content experts did one-on-one training and observed the research nurses collecting the gold-standard mood and psychosis/behavior items. This session also included training on conducting planned in-trainings.

The research nurses returned to their own states, where they trained a facility-nurse from each participating NH on the draft MDS 3.0. We trained one nurse per facility. Nurses could, in turn, elect to train members of their interdisciplinary team (IDT) to complete sections if they desired. We requested that if other team members completed a section, that they do so for at least 20 cases. It appears that some facilities took this approach since many MDS 3.0 forms included multiple entries for times and had titles of different members of the IDT written beside times on the tracking sheet.

Training took place over 3 days at a single location in each state. The training included instruction on scoring all items (both revised and unchanged) in MDS 3.0 as well as on the data collection protocols for the MDS 2.0 and 3.0.

Study Data Collection Design

The study data collection protocols had six purposes.

1. Evaluate MDS 3.0 inter-rater reliability or the extent to which two data collectors achieve the same results when assessing the same resident in the same time frame. It tests the concept's underlying stability and the clarity of the item and instructions. Reliability is a cornerstone for achieving validity when an item is used by different assessors. We measured two types of reliability, research nurse to research nurse, and research nurse to facility-staff. The research to research comparisons evaluated item performance when highly trained nurses used research protocols. The research to facility comparisons measured performance in an operational environment in which one assessor had competing facility responsibilities.

2. Evaluate the feasibility of conducting interviews. To understand the extent to which NH residents could complete the proposed interviews, we examined the percent of interviews attempted and completed rates by all of the residents enrolled in the reliability and facility cross-walk samples.

3. Evaluate MDS 3.0 validity. Validity assesses the degree to which an item measures the intended concept. In addition to literature and expert review, we compared several new items and old items either to established gold-standard assessment items (which are usually longer and more complex) or to similar related items and scales. A comparison to MDS 2.0 required temporally coordinated data collection.

4. Evaluate time to complete. Facility-nurses coded exact start and stop times for each MDS 3.0 case and MDS 2.0 case for all assessment activities. A tracking sheet accommodated possible
interruptions and multiple assessors. Staff were instructed to only include MDS assessment times and not to record time for completing RAPs, team meetings or care planning.

(5) Evaluate staff attitudes about revised assessment. Anonymous feedback was obtained from participating staff through structured surveys. One survey obtained information on normal MDS 2.0 collection processes and baseline attitudes about MDS 2.0. At the conclusion of national testing, we surveyed facility staff and research nurses who participated in the national test to obtain their feedback on MDS 3.0 changes. Respondents were assured that their feedback to both surveys was anonymous.

(6) Maintain MDS 2.0 payment and quality assurance functions.

We designed a comparison of MDS 3.0 to MDS 2.0 items to provide data for developing RUGs adjustment strategies and for mapping items to quality measures. These planned activities required temporally coordinated collection of MDS 2.0 and MDS 3.0 payment and quality items. The related analyses were later modified by CMS and are not reported here.

National Data Collection

Data collection by the research nurses and facility-nurses began September 2006 and continued through February 2007. All MDS 2.0s were collected as part of standard facility protocols and schedules. MDS 3.0 and validation items were timed to coordinate with this schedule. Data were collected on a total of 3822 community NH residents.

Data collectors were instructed to identify cases based on when they were scheduled for MDS 2.0 assessments. The one exception was that the instruction to exclude comatose residents since the associated MDS assessment would be significantly truncated. Part of the training focused on the random assignment of residents who were scheduled for MDS 2.0 assessment into the various protocols designed to meet the study purposes. Online Appendix 4 (available at www.jamda.com) presents in a table format the purpose and data collection approaches for the various protocol types.

One design challenge was ensuring that the MDS 2.0 and 3.0 items for a given resident were collected within a short enough timeframe to allow a fair comparison. This was particularly important for sicker residents, who might undergo significant change over a few days. Therefore both instruments used the same assessment reference date (ARD) and the MDS 3.0 interviews were conducted within 24 hours of the ARD. Nurses were instructed not to view the MDS 2.0 form while collecting the MDS 3.0 and vice versa. The research nurses were instructed to collect the Validation items within 24 hours of the MDS.

We also had to complete the reviews without burdening the resident with multiple proximate interviews. We therefore divided the sample so that residents were assigned among different review types, and therefore all forms were not collected on every resident. When facility MDS 2.0 assessments were due, an algorithm guided data collectors in assigning that case to a review type. In addition, since ideal inter-rater reliability involves coding the same information, the interviewers observed the same interview but each coded independently without discussing observed content or responses. Interviews were alternated between members of each pair. Medical record review was also independently coded.

Analyses of National Test

We created analytic samples to match the study purposes. To assess reliability, we created one analytic data set with research to research nurse assessments and another with research to facility staff assessments on the same resident. We then computed kappa (k) statistics to correct for chance agreement; Pearson correlation coefficients; and intraclass correlations for measures made on a continuous scale. For binary and categorical items, we used unweighted k; for ordinal and scaled items, we used weighted k. We used accepted standards for k: values below 0.4 are considered poor agreement, 0.4–0.6 as moderate, 0.61–0.8 as very good, and those above 0.8 as excellent. Individual item reliabilities for each comparison are available in the appendix to the final project report.61

To compare item distributions between MDS 2.0 and MDS 3.0, we created a feasibility file that included all non-validation cases with a MDS 3.0 form and MDS 2.0 form. We created a different analytic data set with validation criterion items and related MDS 3.0 and MDS 2.0 items. Finally, to ensure that the criterion measures were consistently collected, we assessed the reliability of the research to research nurse collection of criterion measures on a sample of 141 residents.

Feedback survey responses were entered into Excel. The response categories strongly agree and agree were combined as were the categories disagree and strongly disagree.

Results

Analytic Sample Results

Of the 3822 residents enrolled in the study, five cases did not have the full pair of forms needed for their assigned analysis arm and were therefore removed from the analyses. These were cases that were discharged before the paired assessment could be conducted. Of the 3817 remaining, 418 were in the validation arm; 141 were in the sample to check reliability of the criterion measures; and 3258 were in the feasibility sample (2010 from the facility only assessments and 1248 from the reliability samples).

Resident Voice

The majority of residents were able to complete the interview sections. For the feasibility sample of 3258 residents who had MDS 3.0 and MDS 2.0 assessments initiated, 90% completed the cognitive assessment62; 86% completed the mood interview63; 85% completed the interview for preferences for customary routine and for daily activities (families or proxies completed an additional 4%) and 87% completed the pain interview. External auditory assistive devices (“pocket talkers”) supplied by the research team were used to complete interviews with 10% of the sample.

Reliabilities Overall

Overall, the MDS 3.0 items showed either excellent or very good reliability. A few items had only moderate agreement. These included the “other” categories in each diagnoses area where no algorithms or definitions were provided, research nurse assessment of hospice care and report of whether a significant other participated in the assessment. Item level kappa scores for other MDS 3.0 items were very good to excellent for both the research to research nurse and the research nurse to facility comparisons. The kappa values for the two comparisons differed by a categorical level (excellent vs very good) for the MDS 2.0 cognitive performance scale (based on staff assessment and retained for residents unable to complete the BIMS), mobility prior to admission and restraints. MDS 3.0 overall item reliabilities were often higher than those found in studies for related MDS 2.0 items including the new or reformatted items assessing cognition, delirium, mood, hallucinations, delusions, behavior, rejection of care (2.0 item = resists care),
Changes from multiple perspectives. In the development phase, we interviewed and ADL items, were positive. Table 2 shows staff members someone about return to community. Assessment of exudate amount testing that had been problematic in the pilot and revised for national rated problematic. Most involved a few resident interview items (Table 1).

The development and testing of MDS 3.0 evaluated proposed changes from multiple perspectives. In the development phase, we employed an iterative process that incorporated provider and consumer input, expert consultation, scientific advances in clinical knowledge about assessment, and intensive item development by a national VA consortium. This development process allowed the final national testing of MDS 3.0 to include well-developed and tested items. In the national trial in 71 community and 19 VA NHs, the MDS 3.0 draft form showed improved overall reliability, reduced time to collect, and improved staff satisfaction when compared with MDS 2.0.

The inclusion of resident voice items represented one of the most significant changes we tested. Both research nurses and NH staff initially expressed strong reservations about conducting resident interviews. Objections included concerns about resident ability to respond, time demands of interviews, dealing with resident emotions, and lack of skills to conduct structured interviews. After conducting the interviews with actual NH residents, staff feedback shifted dramatically and they wrote many positive comments. One of the more striking comments came from a nurse who said “this reminds me of why I became a nurse.” In addition to changing attitudes, resident interview was feasible. The national MDS 3.0 sample included non-comatose persons scheduled for MDS assessment regardless of cognitive or communication abilities. In this sample, 85% or more completed each interview section.

The national testing of MDS 3.0 provides hope for improving the clinical relevance and efficiency of the assessment. Before the national test, we hypothesized that the new instrument would take longer, because staff would be unfamiliar with the form, charting would not be set up to populate it and all MDS 3.0 assessments were full assessments (without section T). However, analysis of the actual times revealed that average collection times were 45% less for MDS 3.0 than for MDS 2.0. In addition, we had hypothesized that agreement between research nurse and facility staff would be significantly less than that seen between two research nurses. Instead, we found that categorical agreement was similar for facility to research comparisons, indicating a potentially more functional tool for actual NH operations.

Improving the reliability, accuracy, and usefulness of the MDS could have profound implications for improving and measuring the quality of NH care. At the system level, MDS data can inform longitudinal assessments of NH population needs. At the resident level, the MDS is a potentially powerful mechanism to standardize assessment, identifying needs and facilitating care planning and management. However, the full potential of the MDS can be realized only if providers do not view it as an onerous data collection burden and the information is accurate. Although changing established practice is challenging, providers who used the MDS 3.0 provided anonymous feedback that it was more clinically relevant, valid and useful.

Early stakeholder input into MDS offered insight into some of the challenges faced with revising an existing tool linked to quality measurement and payment. Initial feedback suggested the addition of more items than suggested deletions. Many items were used in program function (quality measurement or payment) or had invested advocates or constituencies. In addition, several scales developed by professional organizations were deemed by providers and NH researchers to be too complex and specialized for inclusion in MDS.

### Table 1
**Time to Complete**

<table>
<thead>
<tr>
<th></th>
<th>Average Time (minutes)</th>
<th>Median Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDS 3.0</td>
<td>61.5</td>
<td>60.0</td>
</tr>
<tr>
<td>MDS 2.0</td>
<td>111.6</td>
<td>95.0</td>
</tr>
</tbody>
</table>

ADLs, range of motion, continence, toileting, constipation, several specific diagnoses, pain, end stage disease, oral/dental, and activities. There were no categorical differences between the community reliabilities and the independent VA comparisons of research to research nurse and research nurse to facility-staff agreement on the MDS 3.0 items.

The reliabilities between the research nurses on the collection of the criterion measures for the validation sample were in the excellent range.

**Validation Overall**

For cognition, mood, and behavior items where independent gold-standard measures were obtained for validation, MDS 3.0 showed improved validity. National validation testing for MDS 3.0 cognitive, depression, and behavior items showed significantly higher agreement with criterion measures than did MDS 2.0 items collected on the same residents. Specific validation testing and results for cognition and mood are included in the articles that follow.

**Time to Complete**

The average time for facility staff to complete MDS 3.0 was 45% less than the average time to complete MDS 2.0 on the same sample (Table 1).

**Clinical Relevance**

Staff who used the MDS 3.0 reported that it was more clinically relevant than MDS 2.0 in the anonymous survey. A few items were rated problematic. Most involved a few resident interview items that had been problematic in the pilot and revised for national testing—organized thinking questions, and preference questions for alcohol, participating in activities outside facility and speaking with someone about return to community. Assessment of exudate amount was also rated problematic. Written comments, particularly about the interview and ADL items, were positive. Table 2 shows staff members’ anonymous responses to questions on the MDS 2.0 at the start of trial and Table 3 shows anonymous survey responses at the end of the field trial.

**Discussion**

The development and testing of MDS 3.0 evaluated proposed changes from multiple perspectives. In the development phase, we employed an iterative process that incorporated provider and consumer input, expert consultation, scientific advances in clinical knowledge about assessment, and intensive item development by a national VA consortium. This development process allowed the final national testing of MDS 3.0 to include well-developed and tested items. In the national trial in 71 community and 19 VA NHs, the MDS 3.0 draft form showed improved overall reliability, reduced time to collect, and improved staff satisfaction when compared with MDS 2.0.

The inclusion of resident voice items represented one of the most significant changes we tested. Both research nurses and NH staff initially expressed strong reservations about conducting resident interviews. Objections included concerns about resident ability to respond, time demands of interviews, dealing with resident emotions, and lack of skills to conduct structured interviews. After conducting the interviews with actual NH residents, staff feedback shifted dramatically and they wrote many positive comments. One of the more striking comments came from a nurse who said “this reminds me of why I became a nurse.” In addition to changing attitudes, resident interview was feasible. The national MDS 3.0 sample included non-comatose persons scheduled for MDS assessment regardless of cognitive or communication abilities. In this sample, 85% or more completed each interview section.

The national testing of MDS 3.0 provides hope for improving the clinical relevance and efficiency of the assessment. Before the national test, we hypothesized that the new instrument would take longer, because staff would be unfamiliar with the form, charting would not be set up to populate it and all MDS 3.0 assessments were full assessments (without section T). However, analysis of the actual times revealed that average collection times were 45% less for MDS 3.0 than for MDS 2.0. In addition, we had hypothesized that agreement between research nurse and facility staff would be significantly less than that seen between two research nurses. Instead, we found that categorical agreement was similar for facility to research comparisons, indicating a potentially more functional tool for actual NH operations.

Improving the reliability, accuracy, and usefulness of the MDS could have profound implications for improving and measuring the quality of NH care. At the system level, MDS data can inform longitudinal assessments of NH population needs. At the resident level, the MDS is a potentially powerful mechanism to standardize assessment, identifying needs and facilitating care planning and management. However, the full potential of the MDS can be realized only if providers do not view it as an onerous data collection burden and the information is accurate. Although changing established practice is challenging, providers who used the MDS 3.0 provided anonymous feedback that it was more clinically relevant, valid and useful.

Early stakeholder input into MDS offered insight into some of the challenges faced with revising an existing tool linked to quality measurement and payment. Initial feedback suggested the addition of more items than suggested deletions. Many items were used in program function (quality measurement or payment) or had invested advocates or constituencies. In addition, several scales developed by professional organizations were deemed by providers and NH researchers to be too complex and specialized for inclusion in MDS.

### Table 2
**Overall Feedback on MDS 2.0**

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree and Agree (1–2)</th>
<th>Neutral (3)</th>
<th>Disagree and Strongly Disagree (4–5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDS 2.0 helps NH staff know what is important for assessment.</td>
<td>58% 66%</td>
<td>25% 25%</td>
<td>13% 9%</td>
</tr>
<tr>
<td>MDS 2.0 helps me detect clinical problems that might not have been noticed without the MDS.</td>
<td>55% 45%</td>
<td>26% 28%</td>
<td>19% 28%</td>
</tr>
<tr>
<td>MDS 2.0 items help the NH staff detect changes in the resident that they would otherwise miss.</td>
<td>39% 39%</td>
<td>39% 39%</td>
<td>22% 22%</td>
</tr>
<tr>
<td>MDS 2.0 items fairly reflect the clinical complexity of most residents.</td>
<td>55% 55%</td>
<td>22% 22%</td>
<td>19% 19%</td>
</tr>
<tr>
<td>MDS 2.0 questions are clearly worded.</td>
<td>32% 32%</td>
<td>38% 38%</td>
<td>30% 30%</td>
</tr>
<tr>
<td>MDS 2.0 response choices are clear; choices for specific items are easy to distinguish.</td>
<td>38% 38%</td>
<td>35% 35%</td>
<td>25% 25%</td>
</tr>
</tbody>
</table>
Information from the national study was used to further refine MDS 3.0. In those instances where the proposed MDS 3.0 item performance was no better (or worse) than the MDS 2.0 item, we recommended retaining the MDS 2.0 item, with which facilities had pre-existing experience and training. Participant and expert panel feedback was also used. We also revised the draft instructions used in the national test, incorporating responses to queries from national testing.

The recommended draft MDS 3.0 was further evaluated by various program groups in CMS and CMS made additional modifications. Through this additional process, CMS reinserted some deleted items, increased the look-back period, added some items, replaced the tested ADL items with MDS 2.0 items and modified the goals of stay item. CMS also modified the return to community item.

**Limitations**

The facilities and staff who participated in the study may differ from average facility staff. Although they continued with regular facility responsibilities, they were working in facilities willing to participate in evaluation activities. In addition, they participated in a 3-day multimodal training that we developed and that was taught by trained research nurses. However, the same volunteer facilities provided MDS 2.0 data and 72% of facility MDS coordinators, reported prior formal training in MDS 2.0.

Our national trial was cross-sectional and did not evaluate change over time or repeated assessments. However, many of the new interview items have been shown to be useful, independent of MDS assessment, for serial assessments.

Finally, staff initial hesitancy about conducting interviews might result in some assessors, particularly if not trained or comfortable with using structured interviews, opting out of interviews. Some assessors might assume their residents are unable to respond without first making an appropriate and unbiased effort to optimize the interview. Other assessors may try to bypass the actual interview and complete items without actually completing an interview.

**Future Directions**

Developing a new MDS 3.0 assessment is only an initial step in identifying and responding to the needs of NH residents. Facility response to MDS changes and adherence to interviews should be evaluated over time. Future studies should explore approaches to improving facility care planning and response to findings from the new MDS 3.0 content. Future studies might also evaluate the efficacy of different training approaches and supports for implementing MDS 3.0.

**Conclusions**

Improvements incorporated in the MDS 3.0 national test produced a more efficient assessment: better quality information was obtained in less time. Such gains should improve identification of resident needs and enhance resident-focused care planning. In addition, including items recognized in other care settings is likely to enhance communication among providers. These significant gains reflect the cumulative effect of changes across the tool, including use of more valid items, direct inclusion of resident reports, improved clarity of retained items, briefer assessment periods for clinical items, deletion of poorly performing items, and form redesign.

**Acknowledgments**

The evaluation team, led by RAND and the Harvard Medical School Department of Health Care Policy, benefited from input and engagement by multiple contributors. A list of acknowledgments is included in online Appendix 1 (available at www.jamda.com).

**Supplementary Data**

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.jamda.2012.06.002.

**References**

## Appendix 1

### Participants in VA MDS Pilot Testing

<table>
<thead>
<tr>
<th>Research Group</th>
<th>Investigators</th>
<th>General Area</th>
<th>Specific Topic Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedford VA Center for Health Outcomes Quality and Economics Research</td>
<td>Dan Berlowitz, MD, Elaine Hickey, RN</td>
<td>• Medical Conditions &amp; Complications</td>
<td>• Diagnostic Coding, • Delirium</td>
</tr>
<tr>
<td>Atlanta VA Geriatric Research Education and Clinical Care</td>
<td>Joe Ouslander, MD, Pat Parmelee, PhD</td>
<td>• Geriatric Syndromes</td>
<td>• Pain, • Falls</td>
</tr>
<tr>
<td>Philadelphia VHA &amp; MIRECC</td>
<td>Ira Katz, MD, PhD, Joel Streim, MD, Katy Ruckdeschel, PhD, Suzanne DiFilippo, RN</td>
<td>• Mental Health</td>
<td>• Depression, • Behavior Disorders</td>
</tr>
<tr>
<td>VA Greater Los Angeles Center of Excellence for the Study of Health Care Provider Behavior (Lead Research Team)</td>
<td>PI: Debra Saliba, MD, MPH, Mary Cadogan, GNP, PhD, Josh Chodosh, MD, Karl Lorenz, MD, Maria Edelen, PhD, Patricia Housen, PhD, George Shannon, PhD, Barbara Simon</td>
<td>• Residential life quality, • Mental status</td>
<td>• Customary routine, • Pain and other symptoms, • Goals of care, • Mental status</td>
</tr>
<tr>
<td>Harvard Medical School (Evaluation Team)</td>
<td>Co-PI: Joan Buchanan, PhD</td>
<td>• Evaluation and analysis</td>
<td></td>
</tr>
</tbody>
</table>

D. Saliba, J. Buchanan / JAMDA 13 (2012) 602–610