Published in final edited form as:

Home Health Care Serv Q. 2012 ; 31(4): 267–301. doi:10.1080/01621424.2012.703908.

The Outcome and Assessment Information Set (OASIS): A Review of Validity and Reliability

MELISSA O'CONNOR, PhD, MBA, RN, COS-C and

New Courtland Center for Transitions and Health, School of Nursing, University of Pennsylvania, Philadelphia, Pennsylvania, USA

JOAN K. DAVITT, PhD, MSS, MLSP

School of Social Work, University of Maryland, College Park, Maryland; New Courtland Center for Transitions and Health, School of Nursing, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

The Outcome and Assessment Information Set (OASIS) is the patient-specific, standardized assessment used in Medicare home health care to plan care, determine reimbursement, and measure quality. Since its inception in 1999, there has been debate over the reliability and validity of the OASIS as a research tool and outcome measure. A systematic literature review of English-language articles identified 12 studies published in the last 10 years examining the validity and reliability of the OASIS. Empirical findings indicate the validity and reliability of the OASIS range from low to moderate but vary depending on the item studied. Limitations in the existing research include: nonrepresentative samples; inconsistencies in methods used, items tested, measurement, and statistical procedures; and the changes to the OASIS itself over time. The inconsistencies suggest that these results are tentative at best; additional research is needed to confirm the value of the OASIS for measuring patient outcomes, research, and quality improvement.

Keywords

home care; OASIS (Outcome and Assessment Information Set); reliability; validity

INTRODUCTION

The Outcome and Assessment Information Set (OASIS) is a comprehensive assessment designed to collect information on nearly 100 items related to a home care recipient's demographic information, clinical status, functional status, and service needs (Centers for Medicare and Medicaid Services [CMS], 2009a). The OASIS is completed upon admission,

Address correspondence to Melissa O'Connor, PhD, MBA, RN, COS-C, New Courtland Center for Transitions and Health, School of Nursing, University of Pennsylvania, 418 Curie Boulevard, Philadelphia, PA 19104, USA. omelissa@nursing.upenn.edu.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

discharge, transfer, and change in condition for all Medicare and Medicaid, non-maternity, and non-pediatric beneficiaries. OASIS data are collected by a home care clinician (e.g., nurse or therapist) via direct observation and interview of the care recipient and/or caregiver. Select OASIS indicators are used to assign patients to a Home Health Resource Group (HHRG) for each 60-day home care episode. The HHRG is then used to calculate each patient’s reimbursement rate under the Prospective Payment System (PPS).

The purpose of the OASIS was to provide a standardized assessment tool that would support a case mix adjusted PPS and a mechanism to monitor the quality of care (Davitt & Choi, 2008; Davitt, 2009). A standardized assessment tool was needed which would contain all items essential to measuring a patient’s service needs and quantify that need into a reimbursement level (HHRG; Davitt & Kaye, 2010). Furthermore, a standardized assessment with risk adjustment factors would enable agencies and CMS to monitor performance and modify practice (Shaughnessy, Crisler, Schlenker, & Arnold, 1997a). It is important to consider these purposes when critiquing the reliability and validity of OASIS. Home care clinicians can complete the OASIS to benefit the home care agency in reimbursements or outcome indicators, compromising the reliability and validity of the tool and its value in understanding quality and patient outcomes (Davitt, 2009; Davitt & Choi, 2008; Madigan, Tullai-McGuinness, Fortinsky, 2003). According to CMS, upcoding, or overstating the severity of a patient’s health status, accounted for 11.78% of the change in case-mix between 2000 and 2008 (Davitt & Kaye, 2010; U.S. Government Accountability Office, 2009; Medicare Payment Advisory Commission [MedPAC], 2009; CMS, 2007a, 2007b).

The OASIS was developed over a period of 10 years where data items were refined via multiple research studies (Shaughnessy et al., 1997a, 1997b). Both expert and clinician input and statistical procedures were employed to “measure and risk adjust patient outcomes in home health” (Shaughnessy, Crisler & Schlenker, 1998, p. 64). The research team that developed the OASIS reported interrater reliability kappas ranging from .50 to 1 on functional variables, .6 for dyspnea and pain, and .79 to 1 for behavioral items (Shaughnessy et al., 1994, as cited in Madigan et al., 2003). A second study by the developers of the OASIS found acceptable reliability levels with most items (71%) having weighted kappas of at least .60 (Schlenker, Powell, Goodrich, & Kaehny, 2000, as cited in Hittle et al., 2003). During development of OASIS, a variety of validity analyses were conducted, including: expert consensus validation for outcome measures; criterion-related validity for case mix adjustment items, and practitioner and expert consensus validity on care planning and assessment items (Tullia-McGuinness, Madigan, & Fortinsky, 2009). These early studies were also used to modify the OASIS, for example, changing wording or collapsing response categories (Hittle et al., 2003). Since then, the OASIS has undergone three revisions (four versions of OASIS) to reduce collection time and enhance validity (see Table 1).

Establishment of the validity and reliability of the various OASIS items is of great concern (Hittle et al., 2003; Madigan & Fortinsky, 2004; Kinatukara, Rosati, & Huang, 2005), as is their value as measures of home care quality both individually and as subscales (Sangl, Saliba, Gifford, & Hittle, 2005). These data are not only used to determine reimbursement levels for adequate patient care but to monitor agency-level outcomes. More importantly, data from the various OASIS versions are being utilized currently by researchers to evaluate home care quality, patient outcomes, and to understand the factors which affect quality and/or contribute to disparities in patient outcomes (Brega, Goodrich, Powell, & Grigsby, 2005; Madigan et al., 2003; Peng, Navaie-Waliser, & Feldman, 2003).

Since much of this research is still being conducted on data from earlier OASIS versions, the results of this systematic review have critical implications for research and policy. First, although changes have been made in the OASIS over time, many original variables are still part of the OASIS assessment and their operational definitions have not changed. Likewise, previous versions of the OASIS continue to be analyzed in longitudinal and cross-sectional studies, due to the delay in data availability and the costs associated with purchasing CMS data sets.

To the best of our knowledge this is the first systematic review of the published literature on the psychometric properties of the OASIS since its implementation. The review synthesizes and critiques the existing research on OASIS reliability and validity, focusing on study methods, types of validity/reliability, sampling procedures, items measured, findings, and limitations. Knowing whether the assessment process reliably and accurately captures need is essential to assuring that agencies receive appropriate support to provide quality care. Likewise, understanding accuracy and reliability is essential to monitoring patient stabilization or improvement and agency performance. Finally, outcome and quality research is dependent on valid and reliable measures of key constructs, without which spurious findings may result.

METHODS

A systematic review of the literature was conducted in which PubMed, MEDLINE, CINAHL, Cochrane, and Scopus databases were searched using combinations of the following search terms: Outcome and Assessment Information Set, OASIS, psychometric, validity, and reliability. The reference lists of reviewed articles were also examined for additional studies, specifically those not published in peer-reviewed journals (Berg, 1999). The search was limited to research published in English and conducted in the United States, as the OASIS is utilized in this country only. The search was also limited to studies published after 1999, the year the OASIS was mandated for use.

Twenty-three articles were identified in the search and reviewed. Of these, 11 were eliminated at the abstract review stage, as they did not measure either validity or reliability of the OASIS. Data were extracted from each article in a three-step process. First, articles were identified as evaluating validity, reliability or, in some cases, both. Second, an initial review of the articles was used to develop a standardized narrative review template (Dilworth-Anderson, Williams, & Gibson, 2002; Weiner, Amick, & Lee, 2008). Such a template was required due to the diversity of methods and measures across the included studies. Finally, this template was used to critically analyze each study for types of validity and reliability, methods used, sampling procedures, items measured, significant findings, and limitations. See Tables 2 and 3 for the template and table of evidence (validity studies in Table 2, reliability studies in Table 3). The first author completed a systematic critical review of all articles using the template. The second author then reviewed the articles to validate data extraction. Any disagreement between the authors was discussed until consensus was achieved.

SYSTEMATIC REVIEW RESULTS

The results are divided into two sections, the first addresses studies that measured validity and the second section reports on the reliability studies. A few studies evaluated both validity and reliability; those results are discussed in each respective section.

OASIS Validity

Methods employed—Since 1999, seven studies investigating the validity of the OASIS have been published. Validity is how well the measure captures the concept of interest (Madigan, 2002) or if an item measures what it is intended to measure (Kazdin, 2002). Five of the studies evaluated some form of criterion related validity, the degree to which a measure relates or correlates to some external criterion (Kazdin, 2002; Rubin & Babbie, 2001). Four of these five studies evaluated convergent validity (Trochim, 2006; Weiner et al., 2008), using either a gold standard tool like the Center of Epidemiology Studies Depression Scale (Tullai-McGuinness et al., 2009), Structured Clinical Interview Axis I DSM-IV Disorders (Brown et al., 2004), the certified care plan (Kinatukara et al., 2005), or expert-derived "correct" answers (Madigan et al., 2003) as comparison criteria. Bowles and Cater (2003) evaluated the predictive validity of the case mix weight, the clinical, service, and functional domain scores with regard to risk of hospital readmission. In addition, construct validity was investigated by two studies, using statistical procedures to analyze the relationship between sets of items (Madigan & Fortinksy, 2000) or item response categories (Fortinsky et al., 2003). Fortinsky and colleagues used Rasch modeling and principal components analysis while Madigan and colleagues used principal axis factor analysis.

Studies used a range of designs to assess validity. Although three studies conducted assessments at two time points, the data were treated as cross-sectional in all studies. Four of the seven studies conducted secondary analyses of existing agency clinical records (Bowles & Cater 2003; Fortinsky et al., 2003; Kinatukara et al., 2005; Madigan & Fortinsky, 2000). Two studies employed prospective designs with data collected by agency staff and research clinicians (Brown et al., 2004; Tullai-McGuinness et al., 2009). One study had agency staff conduct assessments on a video simulated case (Madigan et al., 2003).

Studies used different data collection methods. Some studies used only agency staff for data collection (Fortinsky et al., 2003; Kinatukara et al., 2005; Madigan & Fortinsky, 2000; Madigan et al., 2003) while others used a combination of agency and research staff (Bowles & Cater, 2003; Brown et al., 2004; Tullai-McGuinness et al., 2009). Likewise, some studies collected data during the usual agency assessment process while others employed additional assessment instruments administered at different time points. For example, two studies used agency employed clinicians to gather OASIS assessments, but used research staff to complete the comparative instruments (Brown et al., 2004; Tullai-McGuinness et al., 2009). There was variation across studies in terms of professional staff used to conduct the assessments. Brown et al. (2004) utilized registered nurses (RNs) only, while Tullai-

McGuinness et al. (2009) and Madigan et al. (2003) utilized RNs, and therapists. In the two studies that used delayed assessments, the average time lapse between administration of the OASIS assessment and the gold standard instruments ranged from 5 to 23 days.

Training of staff conducting the assessments also varied. Several studies indicated that they assumed that agency clinical staff were adequately trained in OASIS data collection via routine agency training and provided no additional training for the research data collection (Brown et al., 2004; Madigan & Fortinsky, 2000; Madigan et al., 2003). Two studies provided OASIS training for agency clinical staff (Bowles & Cater, 2003; Kinatukara et al., 2005). Two studies reported achieving adequate interrater reliability on the assessments (Bowles & Cater, 2003; Fortinsky et al., 2003). Tullai-McGuinness et al. (2009) provided training to the research RNs administering the comparative instruments; however, they did not assess interrater reliability among the research clinicians. On the other hand, Brown et al. (2004) did not mention research staff training but they did establish excellent interrater reliability ($\kappa = .91$) among the research clinicians conducting the Structured Clinical Interview Axis I DSM-IV Disorders (SCID).

Sampling—Data were drawn from non-representative sampling frames in all studies. Three studies utilized data from between 5 (Tullai-McGuinness et al., 2009) and 29 (Madigan et al., 2003) Ohio-based Medicare-certified home health agencies, while the other studies utilized data from only 1 home care agency. All but two studies (Brown et al., 2004; Kinatukara et al., 2005) used convenience samples. Sample sizes ranged from 141 to 583 patients, 64 to 337 RNs, and 14 to 99 therapists. Four studies investigated the validity of the original OASIS and three investigated the validity of the OASIS-B. No published studies investigating the validity of the OASISB-1 or the OASIS-C were found.

Items/domains measured—The multiple OASIS-items studied fell under several key domains including: functional, clinical, service utilization, behavioral, and affect domains (see Table 2 for specific domains by study). The measurement of validity also differed among the studies. Two of the studies compared the validity of various OASIS items to that of established or "gold standard" tools. The gold standard instruments employed by Tullai-McGuinness and colleagues (2009) included the Short Portable Mental Status Questionnaire (SPMSQ; Pfeiffer, 1975, compared to one OASIS cognitive item), activities of daily living (ADL) and instrumental activities of daily living (IADL) measures of the Older Americans Resource and Services (OARS) Instrument (Fillenbaum & Smyer, 1981, compared to OASIS functional items), the Center of Epidemiology Studies Depression Scale (CES-D; Radloff, 1977), and the Brief Symptom Inventory (BSI; Derogatis & Melisaratos, 1983, compared to one OASIS depression item). The Structured Clinical Interview Axis I DSM-IV Disorders (SCID) was compared to two OASIS depression items (Brown et al., 2004). Although Bowles and Cater (2003) focused on evaluating the predictive ability of the case mix weight, functional, clinical, and service scores of the OASIS with relation to hospital readmission, they also measured the probability of readmission instrument (Pra; Pacala, Boult, Reed, & Aliberi, 1997) as a comparison. Madigan et al. (2003) compared the validity of OASIS HHRG items answered by home care nurses and therapists after video simulation to that of expert opinion.

Statistical procedures used also varied across the studies. To measure construct validity both studies used some form of factor analysis (Fortinsky et al., 2003; Madigan & Fortinsky, 2000). In addition, Fortinsky and colleagues (2003) used Rasch modeling—a type of item response theory that estimates probabilities of item responses to identify measurement challenges within ordinal-level response categories (Hays, Morales, & Reise, 2000). The Rasch model takes into account item difficulty and personal ability and is particularly suited to evaluating sets of variables with nonuniform response items as is the case with the ADL and IADL measures in the OASIS. Several different procedures were used to establish criterion-related validity including: logistic regression (Bowles & Cater, 2003), sensitivity rates (Brown et al., 2004), percent of inconsistencies with the CMS 485 form—the Home Health Certification and Plan of Care required by CMS and signed by the ordering physician (Kinatukara et al., 2005), percent correct answers (Madigan et al., 2003), and Pearson’s correlation (Tullai-McGuinness et al., 2009).

Findings—Construct validity was evaluated by two of the studies. Both found that the functional measures (ADL and IADL) were unidimensional; each measures a single, latent construct (Fortinsky et al., 2003; Madigan & Fortinsky, 2000; Rubin & Babbie, 2001). However, Fortinsky and colleagues (2003) identified two items that were problematic in terms of the item response categories (bathing and telephone use), and that response categories for both measures violated the assumption of equal intervals between response categories. They also found that Rasch modeling improved the accuracy of the OASIS response categories among the functional measures. These findings indicate that the functional measures may “underestimate differences in disability,” especially at extreme values. The authors suggest caution in using these measures to obtain a patient-level disability score and that item response categories require further testing. Madigan and Fortinsky (2000) likewise identified problems in specific domains. They found that the functional domains had high construct validity, but the affect and behavioral domains did not. They recommend that the affect and behavioral domains should be treated as individual items in research or revised by CMS.

Studies concerned with establishing criterion-related validity found the OASIS items that have been tested show low to moderate validity. Correlation between the OASIS cognitive function score and the SPMSQ was .62 (Tullai-McGuinness et al., 2009). Correlation of the OASIS depression item with the BSI was .26 and with the CES-D was .36 (Tullai-McGuinness et al., 2009). Correlation between OASIS ADLs and IADLs and the OARS instrument ranged from .20 to .69 (Tullai-McGuinness et al., 2009). Using the OASIS, nurses identified only 13 of 35 cases with major or minor depression resulting in a 37.1% sensitivity (Brown et al., 2004). Bowles and Cater (2003) reported the Pra was a better predictor of hospital readmission compared to either the case mix weight, or clinical and service scores found on the OASIS. The OASIS functional score performed closest to the Pra in predicting readmission.

Two studies compared OASIS results to those of untested criteria. Madigan et al. (2003) using a video simulation compared RN and therapist responses to expert-derived correct answers. In this study, 58% (11 out of 19) of OASIS items investigated achieved 80% or higher accuracy to the “correct” answer on admission. However, for those eight other items, the accuracy was much lower. Accuracy was even higher for discharge assessments. Madigan and colleagues (2003) report that in many instances where discrepancies were found, nurses were more likely to agree with the correct answer (six out of eight items) than therapists. Discrepancies were small however (maximum of 20%) with more discrepancies noted on the admission OASIS than the discharge OASIS. Kinatukara and associates (2005) compared OASIS data to the certification and care plan forms also completed by agency staff. They found inconsistencies in 48% of the cases between the care plan and OASIS functional items, 26% of cases in the medications category, and 17.7% of cases in prognosis. The lowest proportion of inconsistencies was for enteral feeding.

OASIS Reliability

Methods employed—Reliability captures the consistency of a measure and its ability to generate the same data over repeated applications (Rubin & Babbie, 2001; Weiner et al., 2008). Nine studies evaluated reliability of several OASIS items. Various methods to measure reliability were employed including intrarater reliability (Madigan & Fortinksy, 2000), simulation (Madigan et al., 2003), sequential interrater reliability (Berg, 1999; Hittle et al., 2003; Neal, 2000; Shew, Sanders, Arthur, & Bush, 2010), simultaneous interrater reliability (Madigan & Fortinsky, 2004), both sequential and simultaneous (Kinatukara et al., 2005) and internal consistency (Fortinsky et al., 2003; Madigan & Fortinsky, 2000). In intrarater reliability the same person completes the assessment on the same patient, at two different times, basically providing a measure similar to test-retest reliability (Weiner, et al. 2008). Interrater reliability refers to the degree to which different assessors agree on the item values when assessing the same patient (Kazdin, 2002). In sequential interrater reliability, two clinicians complete the same document at two different times, hours or days apart. Simultaneous interrater reliability uses two clinicians, who independently complete the same assessment at the same time (Madigan & Fortinsky, 2004). Internal consistency reliability tests the amount of agreement or consistency of the items within a domain or scale (Kazdin, 2002). Most of the reliability studies used a prospective cross-sectional design. Two conducted secondary analyses of agency data collected for a larger study (Fortinsky et al., 2003; Madigan & Fortinsky, 2000) and one used an exploratory video simulation (Madigan et al, 2003).

The type of staff used for assessments also varied. Most studies used only agency-employed clinical staff (Berg, 1999; Fortinsky et al., 2003; Madigan & Fortinsky, 2000, 2004; Madigan et al., 2003; Shew et al., 2010), one used only research staff (Hittle et al., 2003), and two used both research and agency staff (Kinatukara et al., 2005; Neal, 2000). Studies measured interrater reliability of various OASIS items between just nurses (Hittle et al., 2003), and between nurses and therapists (Berg, 1999; Fortinsky et al., 2003; Madigan et al., 2003; Neal, 2000; Shew et al., 2010). Three studies did not report the discipline of the involved clinicians but a mix of RNs and therapists is implied (Kinatukara et al., 2005; Madigan & Fortinsky, 2000, 2004). Time between assessments among the five sequential rating studies ranged from 24 (Hittle et al., 2003; Neal, 2000; Shew et al., 2010) to 72 hours (Berg, 1999; Kinatukara et al., 2005).

Procedures to ensure consistent training among data collectors varied. Most studies regardless of whether they employed only clinical or a combination of clinical and research staff to collect data, relied upon the routine agency-based training provided to all staff, assuming this was adequate. Three studies did not discuss how staff were trained (Madigan & Fortinsky, 2000, 2004; Shew et al., 2010). Hittle et al. (2003) and Kinatukara et al. (2005) provided additional training beyond the routine agency-based training for the research staff.

Sampling—The sampling methodologies also differed among the reliability studies. One study used a purposive, quota sampling method (Berg, 1999); four studies used convenience samples; two used random samples (Hittle et al., 2003; Kinatukara et al., 2005); and two did not specify (Fortinsky et al., 2003; Madigan et al., 2003). Several studies used assessment data from only 1 agency (Fortinsky et al., 2003; Kinatukara et al., 2005; Neal, 2000; Shew et al., 2010), one study used 5 agencies (Hittle et al., 2003), while three studies used data from 10–29 agencies located in the same state (Madigan et al., 2003; Madigan & Fortinsky, 2000, 2004) and Berg (1999) included 60 agencies. Sample sizes also varied ranging from 23 to 583 patients and 436 clinicians in the simulation study. Eight studies investigated the reliability of the original OASIS and one investigated the reliability of the OASIS-B1. No published studies investigating the reliability of neither the OASIS-B nor the OASIS-C were found.

Items/domains measured—Three statistical procedures were employed to measure interrater reliability. One study measured simulated interrater reliability reporting response distributions for each discipline and chi-square (similar to percent agreement; Madigan et al., 2003); two studies reported interrater reliability as the percentage of agreement between the raters (Neal, 2000; Shew et al., 2010). Four studies reported both Cohen’s kappa and percent agreement to establish interrater reliability (Berg, 1999; Hittle et al., 2003; Kinatukara et al., 2005; Madigan & Fortinsky, 2004), while Madigan & Fortinsky (2000) used Cohen’s kappa to measure intrarater reliability. In addition, Cronbach’s alpha was employed to measure internal consistency reliability (Madigan & Fortinsky, 2000; Fortinsky et al., 2003).

The items tested for reliability varied greatly between studies. The reliability of between 15 (Fortinsky et al., 2003) and 96 (Hittle et al., 2003) items were investigated. The Home Health Resource Group (HHRG) items, which impact reimbursement, have been among the most commonly studied (four out of nine studies; Berg, 1999; Madigan et al., 2003; Madigan & Fortinsky, 2004; Shew et al., 2010). In most studies, these included specific items from the functional (e.g., ability to ambulate), clinical (e.g., dyspnea), affect, and behavioral domains used to determine the HHRG and thus agency reimbursement. Other studies included the HHRG items but investigated other variables as well (Neal, 2000; Hittle et al., 2003; Kinatukara et al., 2005). One study investigated the overall intrarater and internal consistency reliability of the functional, affect, clinical, and behavioral domains (Madigan & Fortinsky, 2000).

Findings—Internal consistency was high in the functional domain with Cronbach’s alpha ranging from .86 to .91 (Fortinsky et al., 2003; Madigan & Fortinsky, 2000). Internal consistency was low in the affect and behavioral domains as Cronbach’s alpha ranged from .

25 to .56 (Madigan & Fortinsky, 2000). Overall, interrater reliability of the various OASIS items studied ranged from .11 (Kinatukara et al., 2005) to 1.0 (Hittle et al., 2003; Madigan & Fortinsky, 2000, 2004) as measured using Cohen's kappa. Percentage of agreement ranged between 32% (Kinatukara et al., 2005) and 100% (Madigan et al., 2003; Hittle et al., 2003; Madigan & Fortinsky, 2004; Kinatukara et al., 2005) upon admission and between 45.4 and 100% upon discharge (Madigan et al., 2003). In looking at reliability across disciplines, Madigan et al. (2003) found that nurses and therapists agreed on a majority of times at admission (10/16) and discharge (14/16). See Table 4 for itemized list of findings across studies.

Reliability of the HHRG items (identified in Table 4) varied greatly among studies, with some dimensions/items showing high reliability in one study and low or moderate in others. Reliability of HHRG items at admission ranged between 37% (bathing) and 100% (therapies). HHRG items measured with Cohen's kappa at admission ranged from .22 (dyspnea) to .96 (therapy) and at discharge .66 (dyspnea) to 1.0 (multiple HHRG items). Twelve of the 14 HHRG items studied had reliability below .6 or 80% agreement in at least two studies including: vision, pain, presence of open wounds/lesions, dyspnea, bowel incontinence, behavior problems, upper body dressing, lower body dressing, bathing, toileting, ambulation, and transfers. The following assessment dimensions yielded low to moderate reliability on most items: patient medical history, inpatient discharge date, patient prognosis, patient risk factors, supportive assistance, neuro/emotional/behavioral, respiratory issues, elimination status, sensory and instrumental activities of daily living status (average kappas of ≤.60). Patient demographics, living arrangements, and integumentary dimensions (except for the HHRG-item related to open wounds/lesions), resulted in high reliability (kappas ≥.61 on most items). Patient diagnosis variables achieved moderate reliability overall while the Activities of Daily Living dimension showed mixed results across studies and across items. RNs scored higher reimbursement rates in 13 of 24 cases where RNs and PTs disagreed but there was no statistically significant difference between reimbursement rates (Shew et al., 2010).

Measurement of reliability using sequential versus simultaneous ratings also yielded differences. One study that employed both sequential and simultaneous ratings, found that 65% of the items studied (39 items) using sequential ratings had poor interrater reliability (< .40) while only 29% (19 items) had poor interrater reliability using simultaneous ratings (Kinatukara et al., 2005).

General Limitations

These studies clearly have limitations. First, the variability of items studied, methodologies, and statistical procedures employed make comparison between studies difficult. Second, results from these studies may not be generalizable to all Medicare-certified home care agencies as primary sampling frames were non-representative (i.e., either one state or one agency); most used small and convenience samples. For example, one study reported a sample bias of nearly all White subjects and that 98% of the cohort was insured via traditional, fee-for-service Medicare only (Fortinsky et al., 2003). In addition, the lag between the initial assessment and the comparison instrument varied widely across studies.

One validity study reported a delay of up to 5 days on average after the initial OASIS (Tullai-McGuinness et al., 2009) and another study reported a maximum delay of over one month between the assessments (Brown et al., 2004). In the reliability studies the delay ranged from 24 (Hittle et al., 2003; Neal, 2000; Shew et al., 2010) to 72 hours (Berg, 1999; Kinatukara et al., 2005). Such delays present methodological concerns as patient status between the two assessments could have changed, thus effecting either the validity or reliability results of the study. Investigations lack consistency in the items studied and in the methods utilized to study them. Furthermore, these studies evaluated a limited subset of the over 100 OASIS items, addressing key quality improvement domains or areas included in the reimbursement algorithm, but leaving a gap in our understanding of the validity and reliability of the remaining items. Finally, some studies tested aggregate measures (e.g., Bowles & Cater, 2003) while others tested individual items. Thus comparison across studies is not possible. Finally, validity and reliability studies employed a variety of assessors, with some using agency staff and others using research clinicians, some used RNs only and others used both RNs and therapists. The use of different disciplines can impact results due to differing approaches to data collection between PTs and RNs (e.g., observation of patient vs. self-report; Davitt, 2009; Santos-Eggiman, Zobel, & Berod, 1999). For example, Madigan et al. (2003) found significant differences between RN and therapists' completion of OASIS items. Use of research clinicians and simulation may also bias results, in that the added stressors of carrying a caseload are not relevant in a research or simulated situation.

Limitations specific to validity studies—The review discovered several limitations specific to the validity studies. For example, one study reported that the research RNs received a structured orientation and a reference manual but the study investigators did not establish interrater reliability in using the gold standard instruments leaving the validity and reliability of this collected data in question (Tullai-McGuinness et al., 2009). Likewise, interrater reliability was not measured for the agency clinician assessors in most validity studies. Two other studies used untested comparison criteria (expert opinion or patient care plan) to establish convergent validity. Madigan et al. (2003) did not discuss any systematic means to evaluate the "correct" answers derived from expert opinion. Also, use of the care plan to establish validity required subjective review by a research clinician, with no discussion of training for the research staff. This also assumes that the care plan itself is an accurate picture of the patient's status, again without any testing to support that assumption. Likewise, the reviewed studies had inconsistent strategies for assuring that clinical staff assessors were adequately trained in completion of the OASIS. Furthermore, these studies addressed construct and criterion-related validity only. Translational validity (i.e., content validity) was not explored and only one study evaluated the impact of inaccuracies in OASIS completion on reimbursements (Madigan et al., 2003).

Limitations specific to reliability studies—Several limitations in the current OASIS reliability literature have been identified as well. To begin, the methodology of intrarater reliability is not the most effective measure of reliability as the completion of the second OASIS can be influenced by the rater's recall of the initial assessment (Madigan & Fortinsky, 2000). Interrater reliability while more effective, is not without its limitations. Simultaneous raters could compare and change answers or the rater conducting the assessment can influence the information obtained. Madigan and Fortinsky (2004) state that the raters were instructed not to change their answers after discussion with the other rater but there was no discussion of how this was monitored. Sequential rating can pose risks of a patient’s status changing between ratings depending upon how much time has elapsed between assessments, thus affecting reliability. This is an artifact of the measurement process, not necessarily due to poor reliability of the actual items. Other factors that can impact the accuracy of interrater reliability include how raters were trained, differing clinical judgment, and rater expertise and experience. Since few studies controlled for these factors it is difficult to evaluate the potential for error.

DISCUSSION

Overall, the studies included in this analysis indicate low to moderate validity and reliability for some items on the OASIS. However, several studies showed low validity/reliability for certain behavioral, functional, and clinical items, raising concerns regarding their use in outcome measurement either for outcome-based quality improvement (OBQI) or research. For example, construct validity studies demonstrate the unidimensional nature of the functional domain, indicating these items taken together measure a single construct. However, the behavioral and affect domains did not demonstrate clear unidimensionality and thus should be used individually rather than as composite measures. Furthermore, construct validity studies suggest that the response categories may be problematic for certain functional items, and thus aggregate disability scores may not be valid. Criterion-related validity studies suggest mixed findings, with the ADL items generally showing higher validity as compared to “gold standards” or expert opinion and the IADL, affect (e.g., depressed mood) and behavioral items showing much lower validity. Accuracy seems to be better at discharge compared to admission. Similar results were found in the reliability studies. Internal consistency was higher on the functional domain as compared to the affect, behavioral, or clinical domains. Interrater reliability studies, however, showed mixed results even within specific measurement domains (e.g., functional or integumentary status) and across studies. The potential for measurement error in many of these studies is quite high and so these results must be viewed with caution. Given the lack of consistency in methods used, items tested, sampling, statistical procedures, and findings, plus the changes to the OASIS itself over time, these results suggest that additional research is needed.

Research Recommendations

Future research needs to consider issues related to generalizability. This is one of the major gaps in the current knowledge base. Most of the existing studies were conducted with nonrepresentative samples. Studies using multistage probability sampling designs would have greater confidence in any inferences made about the validity and reliability of the OASIS. In addition, while more representative samples may help with the statistical inference by reducing sampling error around validity and reliability estimates, they are not necessarily going to result in better point estimates. Thus, more research is needed to support or refute the findings of the studies reviewed here.

Likewise, consistency in measurement is needed in future research. Researchers employed a number of methodologies in determining validity and reliability of the OASIS. Differences in the gold standard tools used for comparison and the varying measurement methodologies —percentage of agreement, Cohen’s kappa, and Cronbach’s alpha—make comparison across studies difficult. Thus, future research needs to take a consistent approach to measuring reliability and/or validity, using similar statistical procedures, and focusing on key items, in particular those used in OBQI, case mix adjustment, HHRG calculation, and care planning. Specifically, researchers should determine whether individual items are valid and reliable but also whether subscales measuring specific domains (e.g., functional status, affect, behavioral) are valid and reliable. Of concern is whether such scales are unidimensional (Fortinsky et al., 2003; Madigan & Fortinsky, 2000) and whether existing item response categories are valid. For example, Fortinsky and colleagues (2003) have raised important questions about some functional status items and the validity of their response categories. These need to be further evaluated. Also, more careful attention needs to be paid to testing methods and procedures. There are clear drawbacks to both intrarater and interrater reliability, as there are to the use of sequential and simultaneous rating procedures or assessments; however, researchers can employ monitoring and control procedures to address these drawbacks. For example, few studies controlled for assessor experience in home health care or with OASIS, agency factors, or the discipline of the assessor.

Another gap in the literature concerns the type of validity testing conducted. Five studies addressed some type of criterion-related validity, four of which tested convergent, and one tested predictive validity. Discriminant validity, the degree to which a measure does not correlate with a measure of a different construct (Kazdin, 2002), was not explored. Additional research using confirmatory factor analyses of specific OASIS items is recommended. Although construct validity was evaluated by two studies, content validity—the degree to which an instrument captures the range of meanings of the construct—has not been independently tested via expert review (Rubin & Babbie, 2001). It may be time for additional expert review to establish the content and construct validity of specific OASIS items unchanged over time.

In addition, modified items which were implemented with the newest OASIS version are expected to improve validity. This version has added a new evidence-based screening tool, the PHQ-2 (Pfizer, 1999) for depression which is embedded in the OASIS tool but not required as home health clinicians can use other tools for this assessment. The impact of these changes on validity and reliability must be determined. However, it is essential to highlight the problems with these earlier OASIS affect measures, since many research studies are still using data collected via these earlier versions. For example, Tullai-McGuinness et al. (2009) reported inadequate validity for the one depression item, in earlier OASIS versions, as it was not adequately sensitive to identifying depressive symptoms. Approximately 13 to 29% of geriatric patients receiving skilled home care suffer from either major or minor depression, as diagnosed using the Structured Clinical Interview Axis I DSM-IV Disorders (SCID; Bruce et al., 2002; Brown, McAvay, Raue, Moses, & Bruce, 2003; McAvay, Bruce, Raue, & Brown, 2004). Thus, it will be critical to determine in the newest iteration of this measure if the OASIS improves our ability to identify depression.

Likewise, several patient needs assessments which have not yet been independently evaluated for their psychometric properties have been added to replace previous OASIS items (CMS, 2009a, 2009b). Additional items added include: pain, pressure ulcer, risk for hospitalization, influenza and pneumococcal vaccine, heart failure symptoms, fall risk assessments, drug regime review, medication reconciliation, medication education, and a plan of care and intervention synopsis. Since these measures will be used as benchmarks for agency performance and quality improvement as well as in empirical research, we will need to understand whether they improve our capacity to reliably and validly measure key quality improvement indicators in the newest version as agency data are compared from year to year.

Studies also varied in their interpretation of the results. As in other areas of benchmarking, set standards for acceptable levels of reliability and validity should be established via expert review. These benchmarks could then be used to continue to improve the OASIS over time. Other gaps include the lack of investigations related to the accuracy of OASIS completion and empirical evidence to support how to best structure and provide clinician training to assure both OASIS accuracy and reliability. Investigators did not address the home health agency staff’s ability to use the tool or how well home health agency staff were trained in OASIS completion. Differences were found between the disciplines of nursing and therapy (Madigan et al., 2003). Also, discrepancies were found regarding the HHRG OASIS items in both studies (Madigan et al., 2003; Shew et al., 2010) where agencies both over and underestimated the HHRG score. Accurate reimbursement and outcome assessment are important aspects of home care delivery, thus evidence to support the OASIS’ continued use for existing and modified items is critical.

Finally, studies should focus specifically on establishing reliability/validity in a real-world practice context and control for various practice factors. For example, productivity and case management demands can affect an assessor’s performance during the assessment process. In addition, agency incentives to complete the OASIS so as to maximize reimbursement was addressed by one study (Madigan et al., 2003) which found that discrepancies related to the HHRG score did exist but resulted in the home care agency receiving less reimbursement. Despite this study’s findings, agencies could be motivated to make patients appear sicker and more functionally disabled at admission in order to receive higher reimbursements and improve quality scores which could impact the validity of this nationally implemented instrument. This incentive to inflate a patients’ disability must be considered especially when conducting research.

Another significant limit is the lack of controls for training, education, and clinician experience in OASIS completion. Despite the OASIS’ mandatory use for the last decade, neither a standard method of training, nor a minimum competency has been established for home care clinicians or researchers completing the assessments. While CMS has made available an OASIS Guidance Manual and online training modules (http://www.oasistraining.org), no confirmation of usage is required nor any assurance that staff employ the manual and models in a consistent manner.

Longitudinal research using OASIS data sets is problematic; we do not have an adequate number of studies on any one OASIS version to truly establish validity and reliability of individual items or composites. Thus, researchers must exercise caution when using OASIS data to understand outcomes, contributing factors to outcomes, disparities in outcomes, and agency performance/quality.

CONCLUSION

In summary, few studies have published results regarding the validity and reliability of OASIS data since 1999, despite its increasingly common use for home health care policy and research purposes. In addition to being sparse, evidence is inconclusive regarding the validity and reliability of the OASIS and important gaps in knowledge exist. The published research is growing, however, and provides a starting place from which to direct future home care services inquiry. With further research that builds upon current evidence, researchers will be better prepared to test items and conduct more appropriately designed studies to determine the validity and reliability of this data collection tool. Ongoing research on the psychometric properties of the OASIS must be a priority given its role in determining home care reimbursement, home care quality, and its employment in ongoing home health care services research.

Acknowledgments

The authors would like to thank the New Courtland Center for Transitions and Health for their valuable feedback regarding this article. Each author equally contributed to the research and writing of this manuscript.

This work was supported by the John A. Hartford Foundation's Building Academic Geriatric Nursing Capacity Award Program (2010–2012) and the National Institute for Nursing Research (1F31NR012090).

REFERENCES

Berg, K. Interim reliability report: Medicare home health case-mix project. In: Goldberg, HB.; Delargy, D.; Schmitz, RJ.; Moore, T.; Robel, M., editors. Case mix adjustment for a national home health prospective payment system. Second interim report. ABT Associates; Cambridge, MA: 1999. p. G3-G25.

Bowles KH, Cater JR. Screening for risk of rehospitalization from home care: Use of the Outcomes Assessment Information Set and the probability of readmission instrument. Research in Nursing & Health. 2003; 26:118–127. [PubMed: 12652608]

Brega AG, Goodrich GK, Powell MC, Grigsby J. Racial and ethnic disparities in the outcomes of elderly home care recipients. Home Health Care Services Quarterly. 2005; 24(3):1–21. [PubMed: 16203687]

Brown EL, Bruce ML, McAvay GL, Raue PJ, Lachs MS, Nassisi P. Recognition of late-life depression in home care: Accuracy of the Outcome and Assessment Information Set. Journal of the American Geriatrics Society. 2004; 52:995–999. [PubMed: 15161468]

Brown EL, McAvay G, Raue PJ, Moses S, Bruce ML. Recognition of depression among elderly recipients of home care services. Psychiatric Services. 2003; 54(2):208–213. [PubMed: 12556602]

Bruce ML, McAvay GJ, Raue PJ, Brown EL, Myers BS, Keohane DJ, Weber C. Major depression in elderly home care patients. American Journal of Psychiatry. 2002; 159:1367–1374. [PubMed: 12153830]

Centers for Medicare and Medicaid Services. Home health prospective payment system refinement and rate update for calendar year 2008: Final rule. 2007a. Retrieved from http://www.gpo.gov/fdsys/pkg/FR-2007-11-30/pdf/E7-23272.pdf

Centers for Medicare and Medicaid Services. Supporting statement for paper-work reduction act submission–Part A: "Home health quality measures and data analysis.". 2007b. Retrieved from http://www.ilhomecare.org/uploads/pdfs/oasisc.pdf

Centers for Medicare and Medicaid Services. OASIS background. 2009a. Retrieved from http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/OASIS/Background.html

Centers for Medicare and Medicaid Services. OASIS crosswalk. 2009b. Retrieved from https://www.cms.gov/OASIS/Downloads/OASISC3ColumnChangeTable.pdf

Davitt JK. Policy changes in Medicare home health care: Challenges to providing family-centered, community-based care for older adults. Journal of Family Social Work. 2009; 12(4):291–308.

Davitt JK, Choi S. The impact of policy on nursing and allied health services. Lessons from the Medicare home health benefit. Research in Gerontological Nursing. 2008; 1(1):4–13. [PubMed: 20078013]

Davitt JK, Kaye LW. Racial/ethnic disparities in access to Medicare home health care: The disparate impact of policy. Journal of Gerontological Social Work. 2010; 53(7):591–612. [PubMed: 20865622]

Derogatis LR, Melisaratos N. The Brief Symptom Inventory: An introductory report. Psychological Medicine. 1983; 13(3):595–605. [PubMed: 6622612]

Dilworth-Anderson P, Williams IC, Gibson BE. Issues of race, ethnicity and culture in caregiving research: A 20-year review (1980–2000). The Gerontologist. 2002; 42(2):237–272. [PubMed: 11914467]

Fillenbaum GG, Smyer MA. The development, validity, and reliability of the OARS Multidimensional Functional Assessment Questionnaire. The Journal of Gerontology. 1981; 36(4):428–434. [PubMed: 7252074]

Fortinsky RH, Garcia RI, Sheehan JT, Madigan EA, Tullai-McGuinness S. Measuring disability in Medicare home care patients: Application of Rasch modeling to the Outcome and Assessment Information Set. Medical Care. 2003; 41(5):601–615. [PubMed: 12719685]

Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurements in the 21st century. Medical Care. 2000; 38(Suppl. 9):1128–1142.

Hittle DF, Shaughnessy PW, Crisler KS, Powell MC, Richard AA, Conway FW, Engle K. A study of reliability and burden of home health assessment using OASIS. Home Health Services Quarterly. 2003; 22(4):43–63.

Kazdin, AE. Research design in clinical psychology. 4th ed.. Allyn & Bacon; Boston, MA: 2002.

Kinatukara S, Rosati RJ, Huang L. Assessment of OASIS reliability and validity using several methodologoical approaches. Home Health Care Services Quarterly. 2005; 24(3):23–38. [PubMed: 16203688]

Madigan EA. The scientific dimensions of OASIS for home care outcome measurement. Home Healthcare Nurse. 2002; 20(9):579–583. [PubMed: 12352201]

Madigan EA, Fortinsky RH. Additional psychometric evaluation of the Outcomes and Assessment Information Set (OASIS). Home Health Care Services Quarterly. 2000; 18(4):49–62. [PubMed: 11216438]

Madigan EA, Fortinsky RH. Interrater reliability of the Outcomes and Assessment Information Set: Results from the field. The Gerontologist. 2004; 44(5):689–692. [PubMed: 15498844]

Madigan EA, Tullai-McGuinness ST, Fortinsky RH. Accuracy in the Outcomes and Assessment Information Set (OASIS): Results of a video simulation. Research in Nursing & Health. 2003; 26:273–283. [PubMed: 12884416]

McAvay GJ, Bruce ML, Raue PJ, Brown EL. Depression in elderly homecare patients: Patient versus informant reports. Psychological Medicine. 2004; 34:1507–1517. [PubMed: 15724881]

Medicare Payment Advisory Commission. Comments to proposed rule updating PPS rate for 2010. 2009. Retrieved from http://www.medpac.gov/documentss/09012011_2012_HHAPPS_COMMENT_EC.pdf

Neal LJ. OASIS inter-rater reliability. Caring. 2000; 19(8):44–47. [PubMed: 11066980]

Pacala JT, Boult C, Reed RL, Aliberi L. Predictive validity of the Pra instrument among older recipients of managed care. Journal of the American Geriatrics Society. 1997; 45(5):614–617. [PubMed: 9158585]

Peng TR, Navaie-Waliser M, Feldman P. Social support, home health service use, and outcomes among four racial-ethnic groups. The Gerontologist. 2003; 43(4):503–513. [PubMed: 12937329]

Pfeiffer E. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. Journal of the American Geriatrics Society. 1975; 23(10):433–441. [PubMed: 1159263]

Pfizer. The Patient Health Questionnaire-2 (PHQ-2): Overview. 1999. Retrieved from http://www.cqaimh.org/pdf/tool_phq2.pdf

Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1(3):385–401.

Rubin, A.; Babbie, E. Research methods for social work. 7th ed.. Brooks/Cole; Belmont, CA: 2001.

Sangl J, Saliba D, Gifford DR, Hittle DF. Challenges in measuring nursing home and home health quality: Lessons from the First National Healthcare Quality Report. Medical Care. 2005; 43(Suppl. 3):124–132.

Santos-Eggiman B, Zobel F, Berod AC. Functional status of elderly home care users: Do subjects, informal and professional caregivers agree? Journal of Clinical Epidemiology. 1999; 52:181–186. [PubMed: 10210234]

Schlenker, RE.; Powell, MC.; Goodrich, GK.; Kaehny, MM. Quality of home health care: A rural-urban comparison—Final report. Center for Health Services Research; Denver, CO: 2000. Appendix C: Reliability analyses.

Shaughnessy PW, Crisler KS, Schlenker RE. Outcome based quality improvement in home care: The OASIS indicators. Quality Management in Health Care. 1998; 7(1):58–67. [PubMed: 10344983]

Shaughnessy PW, Crisler KS, Schlenker RE, Arnold AG. Outcomes across the care continuum. Medical Care. 1997a; 35(11):NS115–NS123. [PubMed: 9366886]

Shaughnessy PW, Crisler KS, Schlenker RE, Arnold AG. Outcomes across the care continuum. Medical Care. 1997b; 35(12):1225–1226. [PubMed: 9424484]

Shaughnessy, PW.; Schlenker, RE.; Crisler, KS.; Powell, MC.; Hittle, DF.; Kramer, AM., et al. Measuring outcomes of home health care. Center for Health Policy Research and Center for Health Services Research; Denver, CO: 1994.

Shew PA, Sanders SL, Arthur NC, Bush KW. OASIS inter-rater reliability and reimbursement. Home Healthcare Nurse. 2010; 28(1):31–36. [PubMed: 20032729]

Trochim, WMK. Introduction to validity. 2006. Retrieved from http://www.socialresearchmethods.net/kb/introval.php

Tullai-McGuinness S, Madigan EA, Fortinsky RH. Validity testing the Outcomes and Assessment Information Set (OASIS). Home Health Care Services Quarterly. 2009; 28(1):45–57. [PubMed: 19266370]

U.S. Government Accountability Office. Improvements needed to address improper payments in home health (GAO-09-185). Author; Washington, DC: 2009.

Weiner BJ, Amick H, Lee SY. Conceptualization and measurement of organizational readiness for change: A review of the literature in health services research and other fields. Medical Care Research and Review. 2008; 65(4):379–436. [PubMed: 18511812]

TABLE 1

OASIS Version Changes

OASIS version	Year implemented	Significant changes
OASIS	1999	Original document released to all Medicare-certified home health agencies
OASIS-B	2002	Paperwork Reduction Act (Public Law No. 96-511, 94 Stat. 2812)
		Deleted OASIS items not used for payment, quality measurement, or survey purposes in an effort to ease paperwork burden on home care agencies and clinicians
		The deletions made represented a burden reduction of 28%
OASIS-B1	2008	OASIS items added to address clinical domains not currently covered but deemed essential for patient assessment
		Modified item wording or response categories for selected items to reduce the complexity of the tool
		Eliminated seven items not required for payment, quality, or risk adjustment
		Simplified 44 items to promote clarity
		Added 13 process items to support evidenced-based practices
OASIS-C	2010	OASIS-B1 items not used for payment, quality measures (including those used in the survey process), case mix, or risk adjustment purposes (e.g., Transportation and Shopping), were eliminated (CMS, 2009a).
		New items were created to (a) increase clarity in measurement; (b) replace OASIS-B1 items being eliminated; or (c) measure processes of care in home health agencies and to assist clinicians in care planning (CMS, 2009a).

TABLE 2

OASIS Validity Studies

Citation	Type of validity	Method/Design	Sample	Items measured	Findings	Limitations/Gaps*
Bowles & Cater (2003)	Criterion-related Predictive	Evaluated OASIS ability to accurately predict rehospitalization risk, compared to Probability of Readmission (Pra) instrument Secondary analysis of cross-sectional data	147 patients age 65 and over with CHF 1 agency Convenience	HHRG case mix weight, clinical, functional, and service scores Pra instrument Tested OASIS-B	Pra is better at predicting rehospitalization than OASIS case mix weight (HHRG), clinical and service scores, function score was as good at predicting rehosp. as Pra	Small & non-representative sample Limited items studied
Brown, Bruce, McAvay, Raue, Lachs, & Nassisi (2004)	Criterion-related Convergent	Compared agency nurse's ratings of OASIS depressive symptom items to SCID symptom & diagnostic status (blinded) administered by research associates Cross-sectional prospective	220 patients 64 nurses 1 agency Random sample chosen from eligible patients	OASIS "depressed mood" & "diminished interest in most activities" SCID symptom assess & psychologist diagnosis from symptom review Tested original OASIS	OASIS ratings at Adm. did not accurately reflect depression in older home care pts. OASIS depression sensitivity 37.1% (13/35 cases); anhedonia 4.5% (1/22); dep.mood 33% (12/36)	Data from 2/99–12/99 Small & non-representative sample SCID completed from 11 to 48 days after ADM. No discussion of assessor training.
Fortinsky, Garcia, Sheehan, Madigan, & Tullai-McGuinness (2003)	Construct validity Unidimen-sionality & item response categories	Estimated item uni-dimensionality via principal components anal. & identified item response measurement challenges via Rasch modeling Secondary analysis, cross-sectional data	583 patients 1 agency	ADL & IADLs Tested original OASIS	Response categories on bathing & telephone items should be revised; Some items may not accurately capture disability levels; Recommend Rasch modeling for summary disability scores. Unidimensionality supported by PCA	Data collected 11/99–9/00 Non-representative sample (e.g., 96% White, 98% in traditional Medicare) Adm. only
Kinatukara, Rosati, & Huang (2005)	Criterion-related Convergent	Compared OASIS to CMS485 certification & care plan form Used 1 RN hired from agency as research clinician to compare OASIS and 485 Secondary analysis	141 patients 1 agency Random	% of inconsistencies between tools on 13 categories of information found on both tools (e.g., functional, wound, medications, diagnoses, etc.) Tested OASIS-B	Inconsistencies found: Functional, 47.5% Medications, 25.5% Prognosis, 17.7% Diagnosis, 14.2% Wound, 11.3% PT orders, 7.8% Pain, 7.1% Other, 5.7% Shortness of breath, 5.0% Psychosocial, 3.5% Incontinence, 2.1% Senses, 2.1% Enteral, 1.4%	Data from 2002 Non-representative sample No psychometric testing done on CMS485 to confirm its status as a gold standard Different terms used on these tools No control for OASIS assessor Time lapse between completion of OASIS and CMS485

Citation	Type of validity	Method/Design	Sample	Items measured	Findings	Limitations/Gaps*
						No weighting of inconsistencies within each category
Madigan & Fortinsky (2000)	Construct validity	Evaluated how closely related the items are on a specific domain area Secondary analysis from study on outcomes & resource consumption	210 patients 10 Ohio agencies Convenience	Functional, affect, behavioral items Principal axis factor analysis Tested original OASIS	Functional domain items load cleanly & strongly onto 1 factor for adm. & DC Behavioral & Affect domains should be revised or treated as individual items	Non-rep. sample Data from 1996 No control for discipline of the assessor No discussion of staff training on OASIS
Madigan, Tullai-McGuinness, & Fortinsky (2003)	Criterion-related Convergent	Evaluated the accuracy of OASIS completion by agency nurses & therapists to expert derived answers Exploratory simulation	436 clinicians 337 RNs 68 PTs 21 OTs 10 STs Sampling method not reported 29 Ohio agencies	I/ADLs, clinical items, and behavioral items Percent accuracy to expert answers Tested original OASIS	Clinician responses were similar to the "correct" answer for a majority of OASIS items, with more accuracy shown at discharge (76% with >80% accuracy) than at admission (58% with >80% accuracy)	Data from 11/99–9/00 Non-rep. sample No validity tests on expert opinion Not in usual care context No controls for agency factors
Tullai-McGuinness, Madigan, & Fortinsky (2009)	Criterion-related Convergent	Examined criterion validity of key OASIS items to "gold standard" tools Agency staff CRN, PT) completed the OASIS Research RNs completed the gold standard measures within five business days Prospective cross-sectional	203 patients 188 RNs 14 PTs 5 Ohio agencies Convenience	OASIS: I/ADLs, cognitive functioning, & depression Gold standards: I/ADLs of the OARS Instrument, SPMSQ, CES-D Scale, & BSI Pearson's Correlation Tested OASIS-B	ADL item correlations with OARS from .44 to .69; .71 overall IADL from .20 to .68; .49 overall Cognitive with SPMSQ .62 Depressive symptoms .36 (BSI) and .26 (CES-D) OASIS is valid for ADLs & cognition, but may not be sufficiently sensitive for depressive symptoms & IADL items	Data from larger study from 12/99–3/02 but no specific dates reported for this analysis Non-rep. sample Gold standards conducted between 3 & 7 days post OASIS No interrater reliability check on gold standard measures Data collectors received structured orientation & reference manual. No controls for agency factors

[*] Data are listed as reported in original articles.

TABLE 3

OASIS Reliability Studies

Citation	Type of reliability	Method/Design	Sample	Items measured	Findings	Limitations/Gaps*
Berg (1999)	Sequential interrater	Evaluated the reliability of OASIS items for case-mix adjustment purposes; utilized multiple agency staff (RNs, LPNs, therapists) to perform assessments	144 patients 60 agencies 40 staff Purposive quota sample: all patients scheduled to have second visit within week	Pt. demographic, health history, symptoms, therapies, prognoses, risk factors, support, sensory, integumentary, cognitive, behavioral & functional status items Using kappa and % agreement Tested original OASIS	Kappas ranged from poor to moderate on most domains (see table 3)	Data collected 97–98 Up to 3 days between assessments No control for skill levels of assessors Multiple assessors/ agencies without controls
Fortinsky, Garcia, Sheehan, Madigan, & Tullai-McGuinness (2003)	Internal consistency	Evaluated the use of Rasch modeling to improve disability measures Secondary analysis of data from study on home care resource use & outcomes	583 patients 1 agency Patients in this cohort were enrolled into a larger ongoing study Sampling method not reported	I/ADLs via Cronbach’s alpha Tested original OASIS	Internal consistency among all 15 ADL and IADL items in the study cohort was high (Cronbach’s alpha .91)	Data from 11/99–9/00 Non-rep. sample Adm. only
Hittle et al., (2003)	Sequential interrater	Evaluated the reliability of most OASIS items & time burden for OASIS completion Sequential assessments by two research clinicians altering who completed the first and second assessments Prospective cross-sectional study	66 patients (41, Round 1 & 25 Round 2) 5 agencies Random	Pt. demographic, health history, clinical, support, functional, living arrangements Kappa & % agreement Tested original OASIS	Mean kappa = .69 Interrater reliability was excellent (kappa >.80) for many OASIS items and substantial/moderate (kappa >.60) for most items (see Table 3)	Data from 97–98 Not in the usual care context Pt’s assessed at different times but within 24 hrs Nonrep sample No controls for agency factors
Kinatukara, Rosati, & Huang (2005)	Sequential (phase I) & simultaneous (phase II) interrater	Estimated reliability of various OASIS items via comparison of agency & research clinician assessments Phase I, sequential rater (research clinician) visited same patients within 24–72 hours Phase II simultaneous raters (agency & research clinicians)	Phase I, 259 patients 1 agency Convenience Phase II, 105 patients 1 agency Convenience	Various clinical, functional & service items Kappa & % agreement Tested original OASIS	Reliability for many items was considerably lower than prior studies Phase I, 39 items had poor interrater reliability (<.40), 17 had moderate reliability (>.40–60), 2 items had substantial (>.60–80) reliability, and 2 items had excellent reliability (>.80). Phase II, 19 items had poor reliability, 24 items had	Phase I from10/ 00–11/01 Phase II, 01/02–9/02 Nonrep sample Only one research clinician used as rater compared to multiple staff clinicians No control for discipline of agency clinician Phase I time lapse

Citation	Type of reliability	Method/Design	Sample	Items measured	Findings	Limitations/Gaps*
		visit patients jointly Research clinician was a staff nurse with 6 years of home care experience Prospective cross-sectional			moderate reliability, 19 items had substantial reliability, & 4 items had excellent reliability Reliability of 58% of items improved in Phase II	
Madigan & Fortinsky (2000)	Sequential intrarater Internal consistency	Evaluated the reliability of agency clinician assessments within a test-retest framework. Same agency staff completed a second OASIS within 48 hours of the first; the staff person was free to use any clinical data for recall, but was not to refer back to the initial OASIS Secondary analysis	201 patients 10 Ohio agencies Convenience Sampling method not specified	Functional, clinical, affect, behavioral Cronbach's alpha & kappa Tested original OASIS	Internal consistency admission/discharge: Functional—.86/.91 Affect—.36/.56 Behavioral—.24/.53 Intrarater: functional = sufficient; affect & behavioral = good; clinical = low (see Table 3 for detailed kappa results)	Data from 1996 Prior to OASIS mandatory use Intrarater technique used Time between assessments Nonrep. sample No staff training nor controls for assessor discipline; no controls for agency factors
Madigan & Fortinsky (2004)	Simultaneous interrater	Determine reliability of OASIS items based on independent but simultaneous assessments by agency clinicians Prospective cohort	88 patients 21 agencies Convenience	OBQI OASIS items HHRG OASIS items Weighted kappas & % agreement Tested original OASIS	There were no items with values less than .60, and most items had values higher than .70; these findings suggest that the reliability of these OASIS items is sufficient for use in research, regulatory, & reimbursement purposes	Data from 11/99–9/00 Nonrep & small sample Several items had insufficient samples; kappa scores not computed for these items No controls for agency factors
Madigan, Tullai-McGuinness & Fortinsky (2003)	Simulated interrater (by discipline)	Compare agreement between nurses & therapists in OASIS completion via video simulation at adm. & DC Exploratory simulation	436 agency clinicians 337 RNs 68 PTs 21 OTs 10 STs 29 OH agencies Sampling method not specified	Functional items, dyspnea, & pain Response distributions for RNs vs. therapists, χ^2 test of significance Tested original OASIS	Nurses and therapists agreed in their ratings for a majority of items at adm. (10/16) & DC (14/16); the largest significant differences were for: dressing lower body, transferring, & oral med. management	Data from 11/99–9/00 Non-rep. sample Not in usual care context No training provided on OASIS No controls for assessor skill or experience No controls for agency factors
Neal (2000)	Sequential interrater	Measured OASIS reliability between RNs, and between RNs and PTs. 2 nurse research clinicians assessed patients 24 hours after the agency clinician Prospective cohort	23 patients, 14 patients assessed by an RN first; 9 patients by a PT first First 11 patients assessed by Rater 1; last 12 by Rater 2	Functional, affect, behavioral, clinical, demographic items % agreement Tested original OASIS	23 of the 77 items had IRR of .8 or more Poor reliability defined as < .8 found for: reimbursement, financial situation, location of inpt stay, DC date, inpt dx, change in pt's regime after inpt stay, dx requiring change in regime in pt's condition prior to inpt stay, dx &	Data collection dates not specified Non-rep. & very small sample Prior to use of OASIS as part of the clinical assessment No evaluation of research assessors interrater reliability

Citation	Type of reliability	Method/Design	Sample	Items measured	Findings	Limitations/Gaps[*]
			1 agency Convenience sample		severity of, overall prognosis, high-risk factors, structural barriers in the home, safety hazards in the home, informal support, vision, intractable pain, pt has a skin lesion, pressure ulcer, stasis ulcer or a surgical wound, incontinence, cognitive, affect, behavioral, dyspnea, & functional items >RN experience in HH increased reliability; no difference based on experience for PTs	Inferential stats not appropriate due to sample size & bias Unclear description of reliability coefficient and how calculated
Shew, Sanders, Arthur, & Bush (2010)	Sequential interrater	Evaluated reliability between nurses and PTs for assessments conducted within 24 hours RN completed the initial assessment via laptop computer; PT completed a second OASIS on paper Collected within an operational agency using agency staff Prospective cross-sectional	52 patients, 18 RNs, 12 PTs 1 agency Convenience sample	HHRG scores & projected reimbursement rates compared between the two assessments Tested OASIS-B1 Wilcoxon Signed-Rank test	28 cases (54%) had HHRG scores and reimbursement rates that were equal between RNs and PTs Of the nonequal scores: five reimbursement rate projections were separated by less than 10%; 10 reimbursement rate projections differed between 10% & 20%; 9 rate projections differed between 20% & 30% Overall, the RNs scored higher in 13 of 24 nonequal cases No statistically significant difference between the dollar value of RN & PT ratings	Data collection dates not specified Non-rep. & small sample size Adm. only Did not evaluate individual OASIS items, just the overall agreement of HHRG and estimated reimbursement rates Relied on existing agency training of assessors Did not control for assessor experience with OASIS

[*] Dates are listed as reported in original articles.

TABLE 4

OASIS Reliability Findings*

OASIS item: Items are listed in the order in which they appear in the assessment tool	Berg (1999)		Hittle et al. (2003)		Kinatukara et.al. (2005) Sequential/ Simultaneous		Madigan & Fortinsky (2000) Admission/ Discharge	Madigan & Fortinsky (2004)		Neal (2000)
	K#	%!	K#	%!	K#	%!	K#	K#	%!	%!
Gender	1.00	100	1.00							
Race/Ethnicity	1.00	100	1.00			97.7/98.6**				≥80
Payment sources	.47**	94.6**	.70			93.5/95.2				≤79
Reason for assessment	.18	87.2								
Financial situation±	.43**	92.3**				91.2/95.7				≤79
Years of schooling±	.80	95.5								
Primary language±	.87	98.5								
Requires interpreter±	.66	98.5								
Inpatient facility Ddscharge past 14 days	.87	93.4	.52			95/96.8				≤79
Inpatient discharge date		74.5								≤79
Inpatient facility diagnosis		80.2		79						≤79
Medical regime change in past 14 days±	.55	81.1	.78		.12/.12	56/55				≤79
Medical regime change Dx		76.1		74						≤79
Conditions prior to inpatient stay or medical tegime change	.47**	92.4**	.52			84.8/90.1				≤79
Primary Dx$±		76.6		80					90.0	≤79
Primary Dx severity	.43	87.4**	.74							≤79
Other diagnoses		82.7		72						≤79
Other Dx severity	.52	88.7**	.58							≤79
Signs & symptoms±	.41**	91.9**								
Therapy (IV/infusion/ nutrition)$±	.88**	98.9**	.86			99.8/100		.96	99.1	≥80
Adherence to special	.76**	98.5**								

OASIS item: Items are listed in the order in which they appear in the assessment tool	Berg (1999)		Hittle et al. (2003)		Kinatukara et.al. (2005) Sequential/ Simultaneous		Madigan & Fortinsky (2000) Admission/ Discharge	Madigan & Fortinsky (2004)		Neal (2000)
	K$^{\#}$	%$^{!}$	K$^{\#}$	%$^{!}$	K$^{\#}$	%$^{!}$	K$^{\#}$	K$^{\#}$	%$^{!}$	%$^{!}$
treatment schedule$^{\pm}$										
Overall prognosis$^{\pm}$	.50	89.1	.72		.24/.21	78/81				≤79
Rehabilitation prognosis$^{\pm}$	.50	88.7	.77		.27/.11	71/67				≥80
Life expectancy 6 months or less$^{\pm}$	.16	93.8		98	.15/.01	96/96				≥80
High risk factors	.64**	96**	.69			92.8/91.8				≤79
Alcohol/Tobacco use	.44**	94.8**			.34/.44**	95/97**				
Other psychosocial indicators	.53**	97.2**								
Falls frequency$^{\pm}$	.52	94.4								
Structural barriers$^{\pm}$	.35**	87.2**	.52			76.8/80.2				≤79
Safety hazards$^{\pm}$	.59**	97.6**	.56			96.8/96.5				≤79
Sanitation hazards$^{\pm}$	.74**	98.8**	.64			97.1/96				≥80
Current residence$^{\pm}$	.80	96.7	.86		.53/.72	86/87				≥80
Living situation$^{\pm}$	.74**	96.6**	.94			93.2/95.3				≥80
Assisting person	.59**	81.6**	.67			75/78.7				≤79
Primary caregiver$^{\pm}$	.80	94.8	.65							≤79
Frequency of caregiver assistance	.59	89.8	.52		.22/.44	50/48				≤79
Type of caregiver assistance$^{\pm}$	.39**	77.1**	.40			72/74.4				≤79
Extent of help$^{\pm}$	.55**	86.9**								
Caregiver status$^{\pm}$	.56**	94.8**								
Vision$^{\$\pm}$	.53	87.9	.85		.40/.53	78/81		.78	93.6	≤79
Hearing	.52	92.2	.69		.51/.57	76/84				≥80
Speech/Language	.66	93.1	.79		.37/.26	67/78				≥80
Frequency of pain$^{\$\pm}$	.55	83.1	.66		.37/.53	51/58	.61/.74	.77	81.5	≤79
Intractable pain$^{\pm}$	.58	90.1	.67		.37/.35	84/90				≤79

OASIS item: Items are listed in the order in which they appear in the assessment tool	Berg (1999)		Hittle et al. (2003)		Kinatukara et.al. (2005) Sequential/ Simultaneous		Madigan & Fortinsky (2000) Admission/ Discharge	Madigan & Fortinsky (2004)		Neal (2000)
	K#	%!	K#	%!	K#	%!	K#	K#	%!	%!
Presence of open wounds/lesions$	.84	92.4	.85		.37/.56	77/77		.79	90.3	≤79
Presence of a pressure ulcer	.90	95.7	1.00		NA/.69	NA/96	.61/.74			≤79
Number of pressure ulcers by stage	.49**	84.1**	.83		NA/.69	NA/96				≥80
Presence of non-observable pressure ulcer±		84.6		98						≥80
Stage of most problematic pressure ulcer	.86	94.9	.70		NA/.72	NA/94				≥80
Status of most problematic pressure ulcer	.30	76.9	.90							≥80
Presence of a stasis ulcer	.85	97.9	.79		NA/.49	NA/98				≤79
Number of stasis ulcers	1.0	100	1.00							≥80
Presence of non-observable stasis ulcer		100		98						≥80
Status of most problematic stasis ulcer		100	1.00							≥80
Presence of a surgical wound	.95	97.9	.84		NA/.72					≤79
Number of surgical wounds±	.55	91.1	.84							≤79
Presence of non-observable surgical wound±	1.00	100		100						≤79
Status of most problematic surgical wound	.49	83.3	.95		NA/.52					≤79
Dyspnea$±	.49	83.5	.82		.22/.42		.54/.66	.76	64.6	≤79
Respiratory treatments	.77**	98.6**	.95			98.8/100				≥80
Urinary tract Infection in past 14 days	.61	94.2	1.00		.28/.75	80/61		.86	97.5	≥80
Urinary incontinence severity±	.53	81.8	.88		.14/.48	52/88				≤79
Urinary incontinence or catheter presence	.77	93.4	1.00		.20/.81	54/93		.77	90.4	≤79
Bowel incontinence$±	.66	94.8	.73		.39/.67	79/88		.87	95.3	≥80

OASIS item: Items are listed in the order in which they appear in the assessment tool	Berg (1999)		Hittle et al. (2003)		Kinatukara et.al. (2005) Sequential/ Simultaneous		Madigan & Fortinsky (2000) Admission/ Discharge	Madigan & Fortinsky (2004)		Neal (2000)
	K[#]	%[!]	K[#]	%[!]	K[#]	%[!]	K[#]	K[#]	%[!]	%[!]
Ostomy	.85	99.3	.66		1.00/.91	100/99				≥80
Cognitive function	.63	92.9	.63		.33/.47	59/70				≤79
Confusion	.62	90.3	.68		.29/.43	63/70		.66	86.7	≤79
Anxiety	.44	83.5	.61		.11/.46	41/65				≤79
Depression	.30[**]	90.7[**]	.54			85.3/91.3				≤79
Patient behaviors[±]	.29[**]	86.5[**]	.44			80.1/88.7				≤79
Type of behaviors[$±]	.50[**]	95.6[**]	.52		.27/.45[**]	43.5/47.5[**]		.86	97.8	≤79
Frequency of behavior problems	.37	92.8	.96		.21/.33	88/93				≥80
Psychiatric nursing services received[±]		99.3		98						≥80
Hopelessness[±]		89.6	.56				1.0/.84			
Thoughts of death[±]		97.2	.53				1.0/no variance			
Crying Spells[±]							1.0/1.0			
Withdrawal[±]	.31	88.2	.57				1.0/1.0			
Sleep problems[±]	.27	79.2	.59				.69/.84			
Grooming	.63	86.3	.72		.38/.50	48/67	.75/1.00	.81	88.1	≤79
Prior grooming[±]	.62	88.5	.63		.33/.50	53/70				
Dressing, upper[$±]	.68	88	.68		.46/.62	48/66		.89	90.8	≤79
Prior dressing, upper[±]	.70	90.4	.59		.34/.54	51/67				
Dressing, lower[$±]	.71	87.7	.78		.42/.61	50/62		.88	89.6	≤79
Prior dressing, lower[±]	.71	90	.76		.36/.53	51/64				
Bathing[$±]	.68	88.3	.77		.37/.53	37/51	.45/.67	.78	75.8	≤79
Prior bathing[±]	.69	88.4	.67		.25/.38	32/45				
Toileting[$±]	.82	95	.86		.55/.80	65/85		.87	94.1	≤79
Prior toileting	.74	93.1	.65		.31/.70	62/76				

OASIS item: Items are listed in the order in which they appear in the assessment tool	Berg (1999)		Hittle et al. (2003)		Kinatukara et.al. (2005) Sequential/ Simultaneous		Madigan & Fortinsky (2000) Admission/ Discharge	Madigan & Fortinsky (2004)		Neal (2000)
	K$^{\#}$	%$^{!}$	K$^{\#}$	%$^{!}$	K$^{\#}$	%$^{!}$	K$^{\#}$	K$^{\#}$	%$^{!}$	%$^{!}$
Transferring$^{\$\pm}$	.76	94.6	.79		.47/.71	57/75	.58/.87	.72	82.7	≤79
Prior transferring$^{\pm}$	.71	94	.57		.35/.46	54/59				
Ambulation$^{\$\pm}$	.77	94.7	.87		.62/.72	66/76	.72/.90	.85	86.1	≤79
Prior ambulation$^{\pm}$	.72	93.2	.69		.33/.55	48/66				
Eating/Feeding	.48	88.5	.89		.34/. 38	58/65		.67	90.8	≤79
Prior eating$^{\pm}$	.59	91.1	.59		.22/.32	54/64				
Meal prep	.58	81.9	.71		.38/.38	53/53	.41/.75	.81	88.4	≤79
Prior meal prep$^{\pm}$	.69	87.4	.47		.29/.31	48/50				
Transportation$^{\pm}$	.52	95.8	.63		.23/.37	78/83				≤79
Prior transportation$^{\pm}$	.56	88.2	.64		.13/.30	63/71				
Laundry$^{\pm}$	.48	80.9	.64		.18/.59	75/85		.83	87.4	≤79
Prior laundry$^{\pm}$	.59	80.2	.50		.22/48	50/62				
Housekeeping$^{\pm}$	.50	81.1	.54		.30/.44	56/62		.80	80.4	≤79
Prior housekeeping$^{\pm}$	.65	84.1	.46		.24/.31	43/43				
Shopping$^{\pm}$	.50	84.5	.65		.32/.36	55/51		.75	78.0	≤79
Prior shopping$^{\pm}$	.61	83.0	.62		.28/.34	42/40				
Telephone	.71	93.2	.73		.61/.64	71/79		.83	95.3	≥80
Prior telephone$^{\pm}$	.71	94.2	.56		.55/.60	69/75				
Number of meds$^{\pm}$	.80	95.6								
Mgmt. of oral meds	.73	92.1	.82		.54/.73	68/79	1.00/1.00	.91	94.2	≤79
Prior mgmt. of oral meds	.63	91.5	.72		.40/.50	57/67				
Current mgmt of inhalant meds$^{\pm}$	.73	95.4	.91		.54/.58	82/86				≤79
Prior mgmt. of inhalant meds$^{\pm}$	.52	93.5	.91		.49/.42	76/70				
Current mgmt. of injectable meds	.74	94.2	.91		.62/.94	90/97				≤79
Prior mgmt. of injectable	.53	94.0	1.00		.27/.35	73/61				

OASIS item: Items are listed in the order in which they appear in the assessment tool	Berg (1999)		Hittle et al. (2003)		Kinatukara et.al. (2005) Sequential/ Simultaneous		Madigan & Fortinsky (2000) Admission/ Discharge	Madigan & Fortinsky (2004)		Neal (2000)
	K[#]	%[!]	K[#]	%[!]	K[#]	%[!]	K[#]	K[#]	%[!]	%[!]
meds[±]										
Med compliance[±]	.29	83.7								
Patient mgmt of equipment	.67	93.8	.87		054/.66	91/95				≥80
Caregiver mgmt of equipment	.29	79.0		89	.02/.40	40/40				≤79
Therapy need					.06/.60	26/69				
Acute care hospitalization								.84	91	
Discharge to the community								1.00	100	
Overall range by study	.16–1.0	74.5–100	.40–1.00	72–100	.02–1.00/ .01–.94	26–100/ 40–100	.41–1.00/. 66–1.00	.66–1.00	64.6–100	≤79– ≥80

[*] Blank cells indicate the item was not measured in that study.

[$] Indicates the OASIS items that impact the HHRG score.

[±] Indicates the item is no longer found in the OASIS-C.

[!] Percentage agreement 80% minimum acceptance.

[**] Summary value for several items from original study.

[#] Cohen’s kappa (κ): poor IRR ≤.40; moderate IRR .41–.60; substantial IRR .61–.80; excellent IRR ≥.81.