CENTENNIAL VISION

Systematic Review of Studies on Measurement Properties of Instruments for Adults Published in the American Journal of Occupational Therapy, 2009–2013

Hon K. Yuen, Sarah L. Austin

OBJECTIVE. We describe the methodological quality of recent studies on instrument development and testing published in the American Journal of Occupational Therapy (AJOT).

METHOD. We conducted a systematic review using the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist to appraise 48 articles on measurement properties of assessments for adults published in AJOT between 2009 and 2013.

RESULTS. Most studies had adequate methodological quality in design and statistical analysis. Common methodological limitations included that methods used to examine internal consistency were not consistently linked to the theoretical constructs underpinning assessments; participants in some test–retest reliability studies were not stable during the interim period; and in several studies of reliability and convergent validity, sample sizes were inadequate.

CONCLUSION. AJOT’s dissemination of psychometric research evidence has made important contributions to moving the profession toward the American Occupational Therapy Association’s Centennial Vision. This study’s results provide a benchmark by which to evaluate future accomplishments.

In 2007, the American Occupational Therapy Association (AOTA) adopted its Centennial Vision affirming the profession’s goal that occupational therapy be recognized as a science-driven and evidence-based profession. Two years later, when Sharon A. Gutman became the editor-in-chief of the American Journal of Occupational Therapy (AJOT), she named a set of publication priorities designed to support the profession’s progress toward this vision. One of these priorities was to promote the publication of science-driven research evidence related to instrument development and testing (Harries, Gutman, & Polatajko, 2013). After this priority was articulated, the number of articles related to instrument development and testing published in AJOT sharply increased. An electronic literature search in the PubMed database showed that between 2004 and 2008, AJOT published 34 articles (9% of all AJOT articles) related to instrument development and testing. Then, in the 5 yr after Gutman named instrument development and testing as the publication priority (2009 and 2013), AJOT published 84 articles (18% of all AJOT articles) on this topic. This change constituted a 2.5-fold increase over the previous 5-yr period.

Given that research in the area of instrument development and testing in occupational therapy is vital to the survival and growth of the profession (Doucet & Gutman, 2013), it is important to examine not only the number of articles published but also their methodological quality. Hilton, Goloff, Altaras, and Josman (2013) have conducted a comprehensive review of studies...
related to instrument development and testing for children and youth in AJOT between 2009 and 2012. Our study adds to the Centennial Vision series by systematically evaluating the methodological quality of studies on measurement properties of instruments in adult populations published in AJOT between 2009 and 2013.

This systematic review aimed to provide both (1) an evaluation of current scientific evidence on instrument development and testing related to occupational therapy disseminated through AJOT and (2) a benchmark to guide AJOT’s ongoing efforts to publish studies with high methodological quality in the area of measurement over the next 5 yr (i.e., 2014–2018).

Method

Literature Search Strategy

We conducted this systematic review according to the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Moher, Liberati, Tetzlaff, & Altman, 2009). The process began with electronic literature searches performed in the PubMed database for studies on instrument development and testing for adults published in AJOT between January 1, 2009, and December 31, 2013. Queries to identify all relevant articles on measurement properties of instruments were based on Boolean combinations of the following key search terms: (valid* [tiab] OR reliab* [tiab] OR reproducibility of results [MeSH] OR reproducib* [tiab] OR psychometrics [MeSH] OR internal consistency [tiab] OR validation studies [pt] OR coefficient of variation [tiab] OR ceiling effect [tiab] OR observer variation [MeSH] OR discriminative [tiab] OR precision [tiab]). These search terms have been shown to produce results that are both sensitive and precise for finding studies on measurement properties of instruments in PubMed (Terwee, Jansma, Riphagen, & de Vet, 2009).

Study Eligibility Criteria

Articles were included if they met the following criteria: (1) The study was published in AJOT between 2009 and 2013; (2) the article described the development or validation of one or two measurement instruments; and (3) the majority of the study participants were adults (i.e., ≥18 yr old), although children could constitute a minority of the participants in a study. Articles were excluded if (1) they did not contain a study with the specific aim of studying the measurement properties of an instrument; (2) they focused on the assessment process, modification of administering processes or response format, or the clinical utility or usability of an instrument; (3) they were case reports, editorials, or review articles; or (4) the instruments measured the responses of parents of children with special needs (i.e., had a pediatric focus).

Data Extraction and Quality Assessment

The first author (Hon K. Yuen) performed the data extraction. To minimize the risk for selection bias, we both independently screened the titles and abstracts of all the articles that had been retrieved to determine whether they met the eligibility criteria. If there was any doubt, we both performed an additional screening on the basis of the full text of the articles in question. Any disagreements were discussed until consensus was reached. The flow diagram in Figure 1 describes the process used to select articles for this study and the results of this literature search.

We both have doctoral degrees concentrating on measurement, evaluation, and research methodology, and we used the COnsensus-based Standards for the selection of health status Measurement Instruments (COSMIN; Mokkink, Terwee, Patrick, et al., 2010a) checklist to determine the methodological quality of the selected studies. The COSMIN checklist consists of a number of boxes. Each box includes a set of rating criteria (i.e., items) that are relevant to studies of a specific measurement property. We followed the guidelines from the COSMIN manual to select the COSMIN boxes that correspond to the purposes of each study.

Figure 1. Flow diagram of the selection process and study search result.

Note. AJOT = American Journal of Occupational Therapy.
described in each selected article and to score the items in each relevant box.

We both independently appraised the methodological quality of each study of measurement properties in the selected articles. This process was done in two steps. In Step 1, we each independently reviewed the research objectives or questions stated in each selected article to determine which measurement properties (i.e., COSMIN boxes) would be evaluated. In Step 2, we evaluated each article by rating each of the items in the COSMIN boxes that were identified in Step 1. We applied the response options of “yes,” “no,” or “not applicable/indeterminate” to rate each item (Terwee et al., 2012). When our ratings of specific COSMIN items differed, we discussed these items until we reached a consensus. We used an Excel spreadsheet to record our ratings and calculated the percentage of each rating option within each COSMIN box.

Before scoring the selected articles, we studied and discussed both the COSMIN manual and published studies that applied the COSMIN to increase our familiarity with the tool and to clarify our understanding of the terminology and standards of the COSMIN checklist. We also practiced applying the checklist by independently scoring 3 randomly selected articles on measurement properties of instruments for children from AJOT between 2009 and 2013. As a result of this experience, we made some minor modifications to the COSMIN checklist and developed clarifications for some rating criteria. For example, when research staff administered performance-based tests or interviews, data sets were assumed to be complete, and we rated the criterion of “stated percentage of missing items” as “yes”; deletion of participants was acceptable as a description of how missing items were handled; and flaws noted in other COSMIN items were not marked again in the criterion of “important flaws in the design or methods.” Other modifications specific to a particular measurement property are noted in Supplemental Figures 1–6 (available online at http://otjournal.net; navigate to this article and click on “Supplemental”).

COSMIN Checklist

The COSMIN checklist was developed through an international Delphi consensus-based process as a standardized tool for evaluating and quantifying the methodological quality of studies of the measurement properties of health-related patient-reported outcomes (Mokkink, Terwee, Patrick, et al., 2010a). The quality standards described in COSMIN are also relevant to other health-related measurement instruments (Ofstedal et al., 2012). The COSMIN checklist consists of boxes that can be used to rate nine different measurement properties, as follows: internal consistency (Box A, 11 items), reliability (Box B, 14 items), measurement error (Box C, 11 items), content validity (Box D, 5 items), structural validity (Box E, 7 items), hypothesis testing (Box F, 10 items), cross-cultural validity (Box G, 15 items), criterion validity (Box H, 7 items), and responsiveness (Box I, 18 items; Mokkink, Terwee, Patrick, et al., 2010b). The COSMIN boxes for reliability included items that could be used to evaluate studies of test–retest, interrater, intrarater, and other types of reliability. The hypothesis testing box could be applied to studies of convergent or discriminant validity, known-groups validity, and other related methods (Mokkink, Terwee, Patrick, et al., 2010b). Previous studies have demonstrated that the COSMIN checklist has adequate content validity and interrater reliability (Mokkink, Terwee, Gibbons, et al., 2010; Mokkink, Terwee, Patrick, et al., 2010a).

Results

Of the 475 articles published in AJOT between 2009 and 2013, 48 met the eligibility criteria for this study. We evaluated internal consistency in 20 articles, reliability in 20 articles (with test–retest reliability in 5; interrater reliability, including one case of interinstrument reliability, in 15; and intrarater reliability in 5), measurement error in 1, content validity in 6, structural validity in 10, hypothesis testing in 24 (with convergent validity in 11 and known-groups validity in 19), and criterion validity in 13. None of the articles addressed cross-cultural validity.

Two selected articles included psychometric studies that the authors described as examining “sensitivity to change.” These studies were not included in this review because they did not match the COSMIN’s definition of responsiveness (Mokkink, Terwee, Patrick, et al., 2010b).

Table 1 shows which COSMIN boxes (i.e., measurement properties) were assigned to each selected article. Although most studies used classical test theory approaches, 11 studies involved the use of Rasch analysis to evaluate the psychometric properties of the instruments.

COSMIN requires ratings of how authors reported and handled missing data in the criteria for studies on a wide range of measurement properties. In the set of articles included in this study, authors did not always provide the percentage of missing items or describe how they handled missing data or items within a scale. Most other results (described next) were related to COSMIN ratings for specific measurement properties.

Internal Consistency

Thirteen (65%) of the articles assessing internal consistency focused on instruments that were clearly based on measurement models in which all assessment items were theoretically based on a single underlying latent trait. The majority of these authors assessed unidimensionality before evaluating internal consistency, a strong practice that supports the interpretability of internal consistency statistics. Some authors could have strengthened their studies by making clearer links between the theoretical structure of their instruments and the methods that they used to assess the internal consistency of the scales and subscales.

Reliability

According to the COSMIN, a sample size of 50 is the minimum standard for studies that can support reliable conclusions, and in many cases, such as Rasch-based models, much larger samples are desirable (Terwee et al., 2012). Only 3 (15%) of the studies assessing reliability met the minimum criterion of including a sample size of 50, and 50% had sample sizes <30. Sample sizes <30 have a profound negative effect on the
Table 1. Measurement Properties of the Study Articles (N = 48)

<table>
<thead>
<tr>
<th>Study</th>
<th>Item Response Theory</th>
<th>Measurement Error</th>
<th>Content Validity</th>
<th>Structural Validity</th>
<th>Reliability</th>
<th>Hypothesis Testing</th>
<th>Convergent Validity</th>
<th>Criterion Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Be´dard, Parkki, Weaver, Remis, &amp; Dubois (2010)</td>
<td>xx</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canny, Thompson, &amp; Wheeler (2009)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chang, Helfrich, &amp; Coster (2013)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chang, Ailey, Heller, &amp; Chen (2013)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cheng &amp; Cheng (2011)</td>
<td>xx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen, Wang, Crizzle, Winter, &amp; Lanford (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen et al. (2012a)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen et al. (2012b)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen et al. (2011)</td>
<td>xx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen et al. (2012c)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classen et al. (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duquette et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eakman (2012)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engstrand, Krevers, &amp; Kristoffersson (2012)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flin, Pease, &amp; Freimer (2012)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal, Ben, Meir, &amp; Katz (2010)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>George &amp; Croft (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hartman-Maier &amp; Harel &amp; Katz (2009)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hwang (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hwang et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hwang (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jung, Choo, &amp; Lin (2009)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Katz, Bar-Haim Erez, Livni, &amp; Averbuch (2009)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Katz, Averbuch, &amp; Bar-Haim Erez (2012)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kay, Bundy, &amp; Clemson (2009)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>King (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>King, Woodbury, &amp; Velco (2011)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lewis, Forbes, &amp; Tharion (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyons, Li, Kodres, Meenan, &amp; Anales (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meenan, Warren, &amp; Ham (2012)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morrison et al. (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ovsenik et al. (2013)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Study</th>
<th>Item Response Theory</th>
<th>Internal Consistency</th>
<th>Reliability</th>
<th>Measurement Property</th>
<th>Hypothesis Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perlmutter (2013)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rieke &amp; Anderson (2009)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Rowe (2013)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saban, Ornoy, Grotto, &amp; Parush (2012)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Shechtman, Awadzi, Classen, Lanford, &amp; Joo (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Shih, Rogers, Skidmore, Irgang, &amp; Holm (2009)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simmons, Griswold, &amp; Berg (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Søndergaard &amp; Fisher (2012)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stark, Somerville, &amp; Morris (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Su, Tsai, Su, Tang, &amp; Tsai (2011)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toglia &amp; Berg (2013)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsworth, Pallant, Russell, Germano, &amp; Odell (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Weiner, Toglia, &amp; Berg (2012)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong &amp; Moskovitz (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
precision of estimates (Terwee et al., 2012). Also, some authors applied Pearson product–moment or Spearman rank order correlations instead of intraclass correlation coefficients for estimating reliability. A limitation of the Pearson and Spearman correlations is that they do not reflect systematic differences in scores (Lin, 2008).

**Test–Retest Reliability.** The period between the administrations of assessments in studies assessing test–retest reliability varied from 1 wk to 3 mo. These time periods were assumed to be sufficiently long to minimize memory effects; however, in studies with long delays between test and retest administrations (e.g., a 3-mo interval), additional sources of variance may have been introduced. In addition, the interpretability of test–retest reliability results depends on the stability of study participants in relation to the construct to be measured during this interval between assessments. In several studies, certain participants’ characteristics were likely to change between test administrations because of natural recovery or because of interventions received during the interim period. Although instability in participants’ characteristics related to the assessment’s construct can lead to the underestimation of test–retest reliability, the use of participants without clinical conditions may have inflated reliability results in some selected studies (Lucas, Macaskill, Irwig, & Bogduk, 2010). The use of individuals from well populations also affects the interpretability of results because these samples do not reflect the characteristics of clients whom occupational therapists typically see in the clinic.

**Intrarater Reliability.** Most authors clearly stated that raters were blinded to the results produced by other raters. This evidence of independent administration was sufficiently described in most of the studies. Several studies in this set relied on therapists’ ratings of video recordings. Although this methodology can control for some of the challenges inherent in the assessment process, it is likely to overestimate the reliability of assessment tools as they are used in clinical settings. In 1 study, different types of raters assessed participants under different conditions, which may have led to an underestimation of intrarater reliability because both the variability of raters and the variability of assessment conditions may have contributed error to the measurement process.

**Intrarater Reliability.** The majority of the studies assessing intrarater reliability focused on research staff’s consistency in consecutive recordings of participants’ physical parameters. Although establishing this type of consistency is important, it is almost inevitable that the raters in such conditions will introduce expectation bias, which can lead to an underestimation of measurement error. One study was designed to assess convergent validity but mistakenly conceptualized or labeled it as intrarater reliability.

**Content Validity**

All 6 studies on instrument development included processes to assess potential items for their relevance to the construct to be measured and to the target population. Most also evaluated the items on the basis of the purpose of the instrument. The majority of these studies included an evaluation of the degree to which the set of potential items comprehensively represented all relevant aspects of the construct to be measured. Exemplary studies in this category provided clear justification for the number, range, and content of assessment items within a table of specification. In several studies, users were asked to review potential items via cognitive interviewing or similar means, which reflected a strong practice.

**Structural Validity**

The majority of the studies assessing structural validity used approaches to test structural validity that were appropriate to the theoretical structure of the assessment. All 10 studies assessing structural validity included either a factor analysis or a Rasch-based approach to analyze dimensionality. Of the studies in this set, 80% had sample sizes that met the COSMIN’s quality criteria (i.e., samples ≥100 for factor analysis and Rasch models; Terwee et al., 2012).

**Hypothesis Testing**

Hypothesis testing provides critical evidence of construct validity. This type of study involves assessing the degree to which measurement instruments perform in ways that are consistent with hypotheses that are based on the assessment construct (Mokkink, Terwee, Patrick, et al., 2010b). The majority of the studies in this group (i.e., both convergent and known-groups validity) were clearly based on an a priori hypothesis. However, few of these hypotheses specified the expected direction or magnitude of correlations or mean differences. The authors of the COSMIN suggested that when researchers do not clearly and fully articulate their a priori hypothesis, they may be more likely to provide alternative explanations when interpreting low correlations or insignificant results instead of suggesting that the fault lies in the measure (Terwee et al., 2007). Several studies of convergent validity included evaluators who were blinded to the other measures, thus minimizing expectation risk and limiting biased ratings. Only 55% of the studies assessing convergent validity had a sample size ≥50. The statistical methods used in this set of studies consistently met the quality criteria in the COSMIN. However, several authors did not justify their selection of parametric statistics over nonparametric approaches.

**Criterion Validity**

The majority of the studies that assessed criterion validity adopted a reasonable gold standard or best available tool for comparison and conducted adequate statistical analysis including computation of correlations, area under the receiver operating curve, or sensitivity and specificity. In several studies, the raters were not blinded to the results of the gold standard assessment when they administered the outcome assessment or vice versa. Also, in 1 study scores of both the outcome measure and the criterion were derived from the same performance of the participants. This approach can severely inflate the correlation in studies of criterion validity.

**Discussion**

The results of our review indicate that the majority of the selected studies on measurement properties of instruments for adults published in AJOT between 2009 and 2013 had adequate methodological
quality in both design and statistical analysis. Most of these studies focused on hypothesis testing, followed by internal consistency and reliability. Researchers used methodologies based on classical test theory, as well as a range of Rasch-based measurement models, which provided diverse sources of evidence. Specific strengths in methodology included the following: (1) Researchers assessed a wide range of measurement properties; (2) when exploring content validity during assessment development, researchers consistently assessed the relevance and representativeness of potential items; (3) in studies that examined reliability, most researchers stated that raters were blind to other raters’ assessment results; and (4) most studies that focused on hypothesis testing (known-groups validity) and structural validity had adequate sample sizes.

The majority of the authors who addressed missing data used participant deletion as their approach. This approach can lead to lower sample sizes with less statistical power and bias in parameter estimation. Only 1 article in this study described the use of a specific imputation strategy. Wider use of appropriate data imputation strategies could strengthen the quality of future studies. In addition, when researchers used complete data sets without missing data or items within a scale, a clear statement of this fact would remove ambiguity. Another area for improvement is sample size. Of the selected studies, small sample sizes were especially common in studies of reliability and hypothesis testing (convergent validity). Researchers can strengthen their studies by using a priori power calculations.

Future studies related to assessment development could be strengthened by including a description of an assessment specification or blueprint that is clearly linked to the theoretical underpinnings of the assessment’s construct. Defining the construct tested by an assessment, as well as the construct’s structure and intended use, is also a necessary foundation for studies of internal consistency and validity studies based on hypothesis testing. For example, if an assessment includes subscales that are not theoretically unidimensional, then total scores (based on a sum of subscale scores) should not be computed.

The time interval between test administrations is a critical element in the methodology of test–retest reliability studies. Researchers should make an effort to evaluate participants who are in a relatively stable state and use an interim period that is typically used for initial testing and retesting in a particular group of clients in clinical settings (U.S. Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, & Center for Devices and Radiological Health, 2006). Participants in stable states who also have relevant health conditions can add to the interpretability of results. Thus, the variability of participants who experienced a stroke 5–10 yr ago could serve as a baseline when measuring the change of a particular construct after recovery and therapy in people who are experiencing the acute phase of stroke.

In assessing the reliability of assessment scores, researchers should work on developing methodologies that match to clinical assessment conditions. These methodologies could involve (1) including study participants who represent clinical populations, (2) using live observation of performance in studies of interrater reliability, and (3) examining variability that results from administering the same assessment in different (clinical) settings. The use of assessment tools in clinical settings is likely to introduce many sources of measurement error; researchers could use methodologies based on generalizability theory or multifaceted Rasch models to identify the relative contributions of these sources of measurement error in assessment scores when performing one integrated analysis.

It is especially important for studies assessing validity to use both theory and prior research data to develop a priori hypotheses regarding direction and magnitude of associations or mean differences. When applying parametric statistics, researchers should either present evidence that their data meet parametric assumptions or apply alternate nonparametric options.

In the future, the range of measurement properties evaluated in AJOT could be broadened by including more studies that assess measurement error, cross-cultural validation of assessment instruments, and responsiveness studies that evaluate the ability of outcome measures to detect the clinical meaningfulness of changes in scores.

Study Limitations
Although the COSMIN is the most comprehensive tool available for evaluating the quality of studies focused on measurement properties (Mokkink, Terwee, Patrick, et al., 2010b), it only provides rating criteria for a set of traditional psychometric properties. The COSMIN does not provide quality criteria for studies that involve the development of norms scores, nor does it provide criteria for more innovative studies of such issues as the acceptability of assessments to clients; therefore, we did not include these studies in our review. In addition, we based our ratings strictly on information published in the articles and did not contact authors to request details or clarifications. Therefore, some elements that we rated as “unable to determine” may have been quite strong. Finally, the profession’s progress toward the Centennial Vision will be affected by both the quality and the content of occupational therapy assessments; however, the limited scope of this study did not include an examination of the degree to which the constructs represented by these assessments reflect the occupation-based and client-centered values of the profession of occupational therapy.

Conclusion
The dissemination of methodologically strong psychometric research supports the kind of evidence-based decision making that is necessary to enable occupational therapy to be recognized as a science-driven and evidence-based profession, as well as to achieve the broader goals of AOTA’s Centennial Vision. Naming the publication of studies related to instrument development and testing as one of AJOT’s priorities has clearly helped to support this vision. This study provides a benchmark for the evaluation of AOTA’s dissemination of psychometric research in the future. Further achievements in the production and use of psychometric research evidence will require collaboration and communication.

Study Limitations
Although the COSMIN is the most comprehensive tool available for evaluating the quality of studies focused on measurement properties (Mokkink, Terwee, Patrick, et al., 2010b), it only provides rating criteria for a set of traditional psychometric properties. The COSMIN does not provide quality criteria for studies that involve the development of norms scores, nor does it provide criteria for more innovative studies of such issues as the acceptability of assessments to clients; therefore, we did not include these studies in our review. In addition, we based our ratings strictly on information published in the articles and did not contact authors to request details or clarifications. Therefore, some elements that we rated as “unable to determine” may have been quite strong. Finally, the profession’s progress toward the Centennial Vision will be affected by both the quality and the content of occupational therapy assessments; however, the limited scope of this study did not include an examination of the degree to which the constructs represented by these assessments reflect the occupation-based and client-centered values of the profession of occupational therapy.

Conclusion
The dissemination of methodologically strong psychometric research supports the kind of evidence-based decision making that is necessary to enable occupational therapy to be recognized as a science-driven and evidence-based profession, as well as to achieve the broader goals of AOTA’s Centennial Vision. Naming the publication of studies related to instrument development and testing as one of AJOT’s priorities has clearly helped to support this vision. This study provides a benchmark for the evaluation of AOTA’s dissemination of psychometric research in the future. Further achievements in the production and use of psychometric research evidence will require collaboration and communication.
among researchers, theorists, psychometricians, clinicians, journal editors, and peer reviewers. We hope that the results of this study promote discussions of research methodologies in instrument development and testing in these communities of scholars and that these discussions will facilitate AJOT's ability to exceed this benchmark within the next 5 yr. ▲

References


*Indicates studies that were systematically reviewed for this article.


Shechtman, O., Awadzi, K. D., Classen, S., Lanford, D. N., & Joo, Y. (2010). Validity and critical driving errors of on-road assessment for older drivers. American Journal of Occupational Therapy, 64,


