Psychometric Testing of SPIDER: Data Capture Tool for Systematic Literature Reviews

Sherrilene Classen, Sandra Winter, Kezia D. Awadzi, Cynthia W. Garvan, Ellen D. S. Lopez, Swathy Sundaram

KEY WORDS
• evidence-based review
• measurement
• validity and reliability

OBJECTIVE. Systematic literature reviews contribute to evidence-based occupational therapy, yet no data capture tool currently exists to validly and reliably appraise the characteristics and quality of primary studies.

METHOD. We determined the psychometrics of Systematic Process for Investigating and Describing Evidence-Based Research (SPIDER) and piloted it with 201 studies included in a systematic literature review.

RESULTS. Content validity showed item relevance with 73% agreement between two experts. For the quality construct, seven of nine quality indicators were positively ($p < .05$) correlated with the overall quality score. The quality scores were positively correlated ($p < .05$) with two objective measures, inferring criterion validity. Intrarater reliability was moderate to perfect ($κ = 0.4–1.0$). Cross-tab analyses showed less variation in experienced reviewers’ interrater reliability.

CONCLUSION. SPIDER provides plausible opportunities for occupational therapy researchers and graduate students to appraise the characteristics and quality of primary studies but requires testing across other settings.


Occupational therapists require clinical and research competencies to advance practice (Holm, 2000). Evidence-based practice, which includes use of the best available clinical knowledge and reasoning to plan and deliver interventions, may be obtained by conducting a systematic literature review (SLR; Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). An SLR, an exhaustive search of the literature, entails following preestablished protocol to collect, appraise, and synthesize primary studies (original research studies—prospective and retrospective—conducted by a primary investigator) in an unbiased and reproducible way (Cooper & Hedges, 1994). Although SLRs are often associated with determining the effectiveness of interventions, they can also be used to determine underlying etiological (risk and protective) factors (Baker & Tickle-Degnen, 2001; Katrak, Bialocerkowski, Massy-Westropp, Kumar, & Grimmer, 2004; National Health and Medical Research Council, 1999; Towheed, 2005).

Although several data capture tools exist for abstracting information from studies, some content-related limitations are well documented (Zaza et al., 2000). For example, research affirmed that 87% of all data capture tools were specific to the research design, most tools had been developed for experimental designs, huge variability existed among data capture items, few tools had documented evidence of item validity and reliability, and only a few could be used to abstract information from both quantitative and qualitative studies (Katrak et al., 2004).

Context

To best answer an etiological question (determining risk and protective factors associated with older driver safety) asked in a recent SLR (Classen et al., 2006;
Classen & Lopez, 2006), we appraised several data capture tools, including Cochrane Collaboration’s RevMan, Critical Review Form–Quantitative Studies, and tools from the Critical Appraisal Skills Programme (Law et al., 1998a, 1998b; Milton Keynes Primary Care Trust, 2005; Review Manager, 2003). None of the existing tools fully met the need to abstract information for an etiological SLR. For example, some items solely captured open-ended responses (i.e., answers were summarized in narrative format, making scoring difficult) or dichotomous responses (not providing more options to capture detailed responses for improved understanding of the study characteristics). Some tools were not sensitive enough to retrieve information necessary for determining descriptive information (e.g., questions on missing data, diagnostics of models, model fit, or group equivalence were mostly absent) or were unspecific (e.g., not addressing how validity can be determined). Some tools included double-barreled questions (two questions in one statement allowing for only one response; e.g., one item asked, “Have the authors taken account of the confounding factors in the design and the analysis?”). Some tools were exclusive to extraction for quantitative (and not qualitative) studies and failed to capture baseline information (e.g., date of review or keywords). With the exception of RevMan, data capture tools were not embedded in a Web-based electronic data management system.

Measurement Theory for Development of a Data Capture Tool

A data capture tool is, by nature, descriptive because the items capture characteristics of studies and enable a comparison of those characteristics with a quality standard (Kirshner & Guyatt, 1985; Rudman & Hannah, 1998). It is therefore essential that such a tool has a priori, face, content, construct, and criterion validity (Carmines & Zeller, 1979; Streiner, 1993; Walz, Strickland, & Lenz, 1991). The latter assesses how well the items of the tool correlate with another established measure of the same variables, yet no gold standard criterion exists for data capture tools (Katrak et al., 2004; Streiner, 1993). Moreover, given that two or more raters will extract data for an SLR makes establishing inter-rater reliability, the extent of rater agreement between two or more raters, critical. Likewise, the same rater may have to reextract data from the primary sources during different processes of the SLR, making measurement of intrarater reliability, or consistency in rater judgment, equally important (Carmines & Zeller, 1979; Cooper & Hedges, 1994).

Purpose Statement

The purpose of this article is to describe and quantify the psychometric properties of a newly developed data capture tool: Systematic Process for Investigating and Describing Evidence-based Research (SPIDER). Specifically, we report on (1) a priori validity, (2) face validity, (3) content validity, (4) construct validity, (5) criterion validity, (6) interrater reliability, and (7) intrarater reliability. Our aim was to introduce a data capture tool that could be used for future etiological SLRs in occupational therapy and other health-related disciplines (see Appendix).

Methods

Tool Development

The etiological SLR used a mixed-methods approach and necessitated applying a data capture tool suitable to simultaneously, validly, and reliably extract quantitative and qualitative data (Classen & Lopez, 2006). We developed this data capture tool by drawing from various theoretical frameworks, including methodologies used in other systematic reviews (Katrak et al., 2004) and principles of classical test theory (Carmines & Zeller, 1979), measurement theory (Nunnally, 1978; Streiner & Norman, 2003), critical appraisal (Crombie, 2000; Greenhalgh & Taylor, 1997), evidence-based practice (Sackett et al., 1996), qualitative research (Guba & Lincoln, 1989), and SLR guidelines (Cooper & Hedges, 1994).

SPIDER (see Appendix) is primarily a descriptive tool, developed to describe and classify the key characteristics of etiological studies. However, it also contains evaluative properties because the items afford assessment of the quality of the primary sources from which information is extracted. SPIDER is composed of constructs, concepts, domains, and items. Theoretically, SPIDER contains two main constructs (study characteristics and study quality) and four concepts (background [Domain A, 4 items]; screening [Domain B, 7 items]; study descriptors [Domains C–P and R, 129 items]; and quality score [Domain Q, 1 item]). Quality indicators are present in most of the domains. The concepts are divided into 18 domains (e.g., Background–Domain A) with 144 items. The 18 domains, unique to SPIDER, pertain to screening criteria, quality indicators, and an overall quality score. Levels of measurement occur on a multidimensional scale as items yield nominal, ordinal, or discrete numerical data. Several items also capture data in narrative format (e.g., where reviewers are asked to list study objectives: Domain H, Item 1 or 3). Depending on the complexity of the primary study, a trained researcher can abstract a study systematically and comprehensively in approximately 45 to 60 min.

SPIDER went through several revisions, including receiving reviews by two external researchers, evaluation of items by five doctoral-level students enrolled in a measurement
theory class, formal content validity testing, a critical written appraisal by two internationally renowned measurement experts, and reviewer and reliability assessments. In addition, and concurrent with this process, the raters received training and were actively engaged in item development and refinement, and abstracting and pilot testing aspects of this study. This iterative process of assessing and refining the tool took place over a 9-month period.

**Procedure**

We determined validity and reliability of SPIDER, specifically for (1) a priori validity, (2) face validity, (3) content validity using a content validity matrix and a content validity index, (5) construct validity, (6) criterion validity, (7) intra-rater reliability, and (8) inter-rater reliability (Carmines & Zeller, 1979; Streiner, 1993; Waltz et al., 1991; Waltz, Strickland, Lenz, & Soeken, 2005).

**Raters.** For the validity studies, we used two external doctoral-level researchers, affiliated with the University of Florida, to evaluate the face validity; five doctoral students who revised the item selection according to Streiner and Norman’s (2003) measurement criteria; and two internationally renowned measurement experts who each completed a content validity index. For the reliability studies, we used four raters (three doctoral graduate research assistants and one graduate faculty member at the University of Florida) who worked in pairs. All raters had completed basic doctoral-level course work in systematic literature reviews and measurement theory. The raters were involved in the development of the tool from the point of inception. This task included selecting items, formulating and refining the items, pilot testing, rater training, revising the items after feedback, providing feedback to the rest of the team and, consistent with the available literature, developing operational definitions for each of the concepts in the SPIDER tool. With the guidance of the biostatistician, all extracted data were entered into Excel spreadsheets by the research assistants and then exported and analyzed using either SPSS (version 13.0) or SAS (version 9.1).

**Validity.** A priori validity (assurance that necessary constructs are presented in the tool) was established by reviewing existing data capture tools (Law et al., 1998a, 1998b; Milton Keynes Primary Care Trust, 2005; Review Manager, 2003), using principles of SLR (Cooper & Hedges, 1994), and applying appraisal guidelines for primary quantitative and qualitative research studies (Crombie, 2000; Greenhalgh & Taylor, 1997). We ascertained face validity by asking two experienced qualitative researchers to evaluate the tool in terms of its content, orientation, and purpose (Streiner, 1993). On the basis of their feedback, we were able to further refine the tool’s appearance, organization, item construction, and relevance. We constructed a content validity matrix to visually illustrate the assignment of each item to one or more domains. By doing so, we determined the adequacy and extent of items to domain representation.

In addition, two internationally renowned measurement experts completed, individually, a content validity index (CVI) on SPIDER items as a means to rate the tool on content, purpose, organization, clarity, appearance, conciseness, and adequacy. The CVI involved using a 4-point Likert scale to rate the items with regard to their relevance (1 = not relevant; 4 = very relevant). To calculate the CVI score, the scale was collapsed into the dichotomous categories: not relevant (score of 1 or 2) and relevant (score of 3 or 4). Items receiving relevant ratings by both reviewers were tallied and expressed as a ratio (tallied responses divided by total number of items). In addition, these reviewers provided written comments that helped with revisions of the items. We also determined inter-rater reliability between these two experts.

A unique contribution of SPIDER is its capacity to assess study quality. Throughout the instrument there are quality indicators (QI) or questions that can be answered objectively for each extracted source. For example, SPIDER includes a QI that addresses sampling method (Domain D) in the study, and the corresponding item (D2) asks to state the type of sampling method (random, convenience, or other) that has been used. A conceptual framework for the quality construct is displayed in Figure 1. The 15 quality indicators were classified into 9 quality themes (e.g., sampling and participation, statistical analysis, outliers, missing data). At the end of the instrument, the reviewer is asked to assign an overall quality score to the study that ranges from 1 (very poor) to 10 (excellent). In addition, an overall summative quality value was statistically derived by adding and summing the scores of the 15 quality indicators. Regardless of the number of options, a score of 1 was assigned when evidence existed that the item was documented, and 0 was assigned in the absence of evidence. The items and scoring system were not weighted.

To start determining construct validity, we tested the relationships between the 15 individual objective QIs (15 items; see Figure 1) and the reviewer-assigned quality score (based on 9 quality themes; see Figure 1) using Wilcoxon rank sum and Kruskal-Wallis tests (Dawson & Trapp, 1994). In addition, we performed Spearman correlation analysis between the reviewer-assigned quality scores and the overall summative quality values.

Currently, no gold standard exists to determine study quality (Katrak et al., 2004). Although reviewers assigned an overall quality score to each study, we recognized the subjective nature of these ratings and the likelihood that they were influenced by the raters’ level of education, experience,
complexity of the studies, or rater burden experienced during the review. In response, we adopted an objective measure to test criterion validity. Although not without limitations (e.g., many weak studies may be cited), we searched the number of citations listed for each study on the Google Scholar and Web of Science search engines. Citation information was available for 181 studies on Google Scholar and 145 studies on Web of Science. To control for number of years in print, we divided the number of citations by number of years in print to form a citation index (quantitative value) for each search engine. To examine criterion (concurrent) validity and by using Spearman correlations from SAS software (version 9.1, Cary, NC), we quantified the relationship among the SPIDER quality scores (reviewer-assigned quality score and overall summative quality value) and the citation indexes.

Reliability. We assessed interrater reliability of the SPIDER tool during two stages of the SLR: abstract screening (for inclusion of articles in the SLR) and data extraction. We used the functions of MAPLE software version 10.0 (Waterloo Maple Inc., Waterloo, Ontario) to ascertain a random selection of studies for these reliability tests. During

---

**Figure 1. Conceptual framework for the quality constructs representing the 9 quality themes and the 15 objective quality indicators (refer to Appendix A for all the numbered items).**

<table>
<thead>
<tr>
<th>QUALITY CONSTRUCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>9 QUALITY THEMES</strong></td>
</tr>
<tr>
<td><strong>1. SAMPLING AND PARTICIPATION</strong></td>
</tr>
<tr>
<td><strong>2. STATISTICAL ANALYSIS</strong></td>
</tr>
<tr>
<td><strong>3. OUTLIERS AND MISSING DATA</strong></td>
</tr>
<tr>
<td><strong>4. DIAGNOSTICS</strong></td>
</tr>
<tr>
<td><strong>5. MODEL FIT</strong></td>
</tr>
<tr>
<td><strong>6. AUTHOR LIMITATIONS</strong></td>
</tr>
<tr>
<td><strong>7. VALIDITY</strong></td>
</tr>
<tr>
<td><strong>8. RELIABILITY</strong></td>
</tr>
<tr>
<td><strong>9. RATIONALE</strong></td>
</tr>
<tr>
<td><strong>15 OBJECTIVE QUALITY INDICATORS (ITEMS)</strong></td>
</tr>
<tr>
<td>(D1, D2a) Sampling method; (F1) Sample size justification; (D4) Informed consent; (G2b) Loss to follow-up</td>
</tr>
<tr>
<td>(F2a) Quantitative design; (G6) Group equivalence</td>
</tr>
<tr>
<td>(G3) Detection of outliers; (G2a) Item nonresponse</td>
</tr>
<tr>
<td>(G4a) Model diagnostics</td>
</tr>
<tr>
<td>(G5) Model fit</td>
</tr>
<tr>
<td>(N2) Limitations</td>
</tr>
<tr>
<td>(J13f) Validity discussed</td>
</tr>
<tr>
<td>(E1a and E2a) Intra- and interrater reliability; (J13g) Reliability discussed</td>
</tr>
<tr>
<td>(J13h) Rationale discussed</td>
</tr>
</tbody>
</table>

Overall Summative Quality Value (Statistically derived summation of the scores of the 15 quality indicators)
the initial abstract screening stage, we evaluated the level of agreement between the scores of two rater pairs (4 raters consisting of trained graduate research assistants and faculty) on 5% (130/2,509) of the abstracts. We assessed agreement using kappa with the scale quantifying agreement levels beyond chance as follows: 0.0 to 0.2 (none–slight), 0.2 to 0.4 (fair), 0.4 to 0.6 (moderate), 0.6 to 0.8 (substantial), and 0.8 to 1.0 (near perfect–perfect; McGinn et al., 2004).

During the data extraction stage, we selected 20% of the (40/201) studies included in the final SLR. For each of the rater pairs described previously, we measured agreement based on raters agreeing on 33 items representing all the SPIDER domains. The rating for these items yielded four categories of interrater agreement 0 to 2 (poor), 3 to 5 (fair), 6 to 8 (moderate), and 9 to 10 (excellent).

For intrarater reliability, each of the four reviewers reextracted data from 13 randomly selected studies (n = 52; MAPLE, 2005). Although a memory or learning effect was possible, we attempted to reduce rater bias by blinding raters to the study titles and authors. These data were coded, analyzed, and compared with the same reviewers’ first extraction that occurred 3 to 6 months earlier.

Results

Validity

The literature on SLR, critical appraisal of studies, quantitative and qualitative research, and existing instruments yielded numerous topical areas that we categorized into 18 distinct domains (e.g., study design, analysis techniques).

Face Validity. Evaluating these 18 domains and items, two objective reviewers agreed that the data capture tool demonstrated face validity consistent with the tool’s stated purpose, that is, to assist in the critical appraisal of studies and to facilitate data capture (extraction) from studies included in the SLR. They judged the tool as having an appropriate length and breadth of items and that the tool was clearly organized. However, they also provided helpful feedback related to item construction (e.g., double-barreled or unclear items) that enabled us to revise, refine, and clarify the items. Reviewers judged the relevance of the items as acceptable for a critical appraisal or data capture tool.

Content Validity Matrix. Findings from the content validity matrix confirmed that the SPIDER items represented an adequate sampling of items and targeted the appropriate domains.

CVI. Results of the CVI indicated that the two reviewers deemed more than half of the tool’s items to be relevant. This agreement between reviewers translated into a CVI score of 0.60, representing a moderate level of relevance (Waltz et al., 2005). Specific feedback provided by these reviewers enabled us to further revise, delete, or clarify items. A high moderate rate of agreement, 73%, existed between the ratings of the two measurement experts.

Construct Validity. Table 1 gives the mean and standard deviation reviewer-assigned quality score for each of the 15 QIs and the corresponding p value from the test of significance between them. Seven of the QIs were found to be statistically significantly related to reviewer-assigned quality score at the .05 level of significance. The reviewer-assigned quality score and the overall summative quality value were positively and statistically significantly correlated (r = .40, p < .01), providing early support for construct validity.

Criterion (Concurrent) Validity. The Spearman correlation coefficients between the overall quality rating score were fair, yet statistically significantly and positively related to the external criteria: the citation indexes from Google Scholar (r = .31, p > .01) and Web of Science (r = .26, p > .01).

Reliability

Intrarater Reliability. For abstract screening, kappa coefficients ranged from 0.4 to 0.6 for the first pair of reviewers and 0.8 to 1.0 for the second pair of reviewers. For the data extraction process, kappa statistics ranged from 0.2 to 0.8 for the first pair of reviewers and from 0.4 to 1.0 for the second pair of reviewers. After these pilot data were obtained, raters from the first pair of reviewers were retrained until a minimum value of 0.4 was achieved (moderate level of agreement beyond chance), as an adequate indicator for rater reliability (McGinn et al., 2004).

Intrarater Reliability. To determine intrarater reliability, we performed cross-tab analyses between the first and the second rating, for each of the four reviewers, using 48 (of the 52) randomly selected studies that had no missing data. This cross-tab analysis had two functions: First, we determined the percentage of intrarater agreement for each reviewer, by SPIDER item. We made two main observations.

First, reviewers had different levels of variation in percentage agreement between their first and the second extraction. For example, Reviewer 1 showed 92% agreement between the ratings of the first and second extraction for the variable purpose, whereas Reviewer 4 showed a 42% agreement between the first and second extractions of the same variable.

Second, consistency varied within reviewers on rating the same item (e.g., gender) where little interpretation is expected. After examining these obvious inconsistencies, we found rater discrepancies in targeting the gender of the specific group (e.g., main sample vs. subgroups). To correct these inconsistencies, we improved on the operational definitions of the problematic variables and clarified items with dubious meanings. For example, to correct for the gender
intrarater reliability scores for each reviewer by percentage of reliability achieved. This table illustrates that 73% of the total ratings for Reviewer 1 were in the good–excellent category, whereas 55% of Reviewer 2’s total rating, 76% of Reviewer 3’s total rating, and 67% of Reviewer 4’s total ratings were in the good–excellent category.

Discussion

The purpose of this article was to describe the validity and the reliability of the SPIDER data capture tool, developed as an objective means to describe the characteristics of primary studies and to evaluate the quality of those studies included in etiological SLRs.

Validity

A priori validity was established by developing SPIDER to be consistent with the criteria, strategies, and guidelines from specific literature addressing evidence-based practice, critical appraisal, SLR, and qualitative research strategies. Face validity was improved (e.g., refine the purpose, comprehensiveness, item construction, relevance of the tool) on the basis of revisions that addressed feedback from two experienced researchers. Subsequently, all of the tool characteristics that were considered during face validity establishment were reevaluated by means of the content validity matrix. This matrix provided support for adequate representation of items to domains. The CVI score revealed that although further tool refinement was indicated, the items were relevant, and we obtained agreement (moderately high) between the two expert reviewers.

Construct and Criterion Validity

The most important construct underlying this tool, from a conceptual standpoint, is its ability to measure quality. We chose 15 QIs and a subjective measure of overall quality to derive a quality construct. We found that 7 of the 15 quality indicators were significantly related to the reviewer-assigned quality score, suggesting that those indicators meaningfully explained a portion of the variance contained in the quality score. The reviewer-assigned quality scores and overall summative quality values were significantly correlated. This

Table 1. Total Subjective Quality Score by Quality Indicators

<table>
<thead>
<tr>
<th>Quality Indicators</th>
<th>M</th>
<th>SD</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sampling and participation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>7.6</td>
<td>1.4</td>
<td>.06</td>
</tr>
<tr>
<td>Not random</td>
<td>7.0</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>Sample size justification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given</td>
<td>7.6</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>Not given</td>
<td>7.1</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Informed consent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>7.5</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.1</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Loss to follow-up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>7.3</td>
<td>2.3</td>
<td>.54</td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.2</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>2. Statistical analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>6.8</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Bivariate</td>
<td>7.0</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Multivariate</td>
<td>7.6</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Equivalence of groups</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Established</td>
<td>7.6</td>
<td>1.5</td>
<td>.37</td>
</tr>
<tr>
<td>Not established</td>
<td>7.1</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>3. Outliers and missing data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outlier</td>
<td></td>
<td></td>
<td>.87</td>
</tr>
<tr>
<td>Discussed</td>
<td>7.4</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.2</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>Item nonresponse</td>
<td></td>
<td></td>
<td>&lt;.01*</td>
</tr>
<tr>
<td>Discussed</td>
<td>7.9</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.0</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>4. Diagnostics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>7.9</td>
<td>1.4</td>
<td>.02*</td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.0</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>5. Model fit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>8.0</td>
<td>1.5</td>
<td>.01*</td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.0</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>6. Author limitations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>7.5</td>
<td>1.7</td>
<td>&lt;.01*</td>
</tr>
<tr>
<td>Not discussed</td>
<td>6.8</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>7. Validity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>7.9</td>
<td>1.4</td>
<td>&lt;.01*</td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.0</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>8. Reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>7.9</td>
<td>1.3</td>
<td>&lt;.01*</td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.0</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Rater reliability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Determined</td>
<td>7.8</td>
<td>1.7</td>
<td>.09</td>
</tr>
<tr>
<td>Not determined</td>
<td>7.1</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>9. Rationale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed</td>
<td>7.2</td>
<td>1.8</td>
<td>.72</td>
</tr>
<tr>
<td>Not discussed</td>
<td>7.1</td>
<td>1.5</td>
<td></td>
</tr>
</tbody>
</table>

*Result from Wilcoxon rank sum test (for indicators with two levels) or Kruskal-Wallis test (for indicators with three levels).
*p < .05.

error, we amended Item 10 in Domain J to read “socio-demographic variable by whole sample or subgroups” instead of “socio-demographic variable by sample.”

Second, to make meaning of this cross-tab analysis, we recoded each reviewer’s percentage of agreement between the two ratings as less than 50% (poor–fair), 50% to 74% (moderate–good), or 75% to 100% (good–excellent; Portney & Watkins, 2000, p. 65). Table 2 presents the recoded
The American Journal of Occupational Therapy

Future Development

Ongoing development of SPIDER pertains to standardizing the tool and developing the Web-based version with definition hyperlink capabilities for easy access to conceptual definitions, enhancing programming functions for reporting purposes, building an algorithm for assigning the quality score, and translating it into Latin American languages. We are also completing guidelines to use this tool as an educational resource in teaching critical appraisal and data extraction concepts to future generations of health professionals.

Summary

With the increased demand for evidence-based practice and research, occupational therapy scholars must develop, test, and document the empirical methods used to summarize the growing body of literature. In this study, we have introduced SPIDER, a new data capture tool, developed to extract study characteristics and to provide a quality score of those studies included in etiological SLRs. We have tested the utilities of this tool by abstracting information from studies included in a completed SLR and by producing research reports used for analysis and synopsis of findings (Classen et al., 2006). We have reported on the basic psychometric properties of this tool and provided suggestions to improve the utility of the tool in developing a Web-based version. This instrument in its current format provides plausible opportunities for occupational therapy (and other) researchers and graduate students to select, appraise, analyze, and report on bodies of literature. As such, occupational therapy (and other) researchers can use this tool in comprehensive or systematic literature reviews, and graduate students trained in measurement theory may benefit from using this tool to summarize literature for dissertation or thesis-based research.

Acknowledgments

We thank Chris Faircloth for his reviews and valuable feedback; Nita Ferree (reference librarian) for her assistance; the five doctoral students who critiqued the items; and Jim Hinojosa and Mary Law, who shared their expertise as raters for the content validity index study. This research is funded by the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention sponsor award #K01 CE000497 (2004–2007). The principal investigator is Sherrilene Classen, Department of Occupational Therapy, College of Public Health and Health Professions, University of Florida.

Reliability

The tool was sufficiently reliable among rater pairs in identifying suitable abstracts for the SLR. Nevertheless, for data extraction, a very complex process, we observed poor to moderate reliability among novice researchers and moderate to good reliability for more experienced researchers. Given that in any SLR team members will vary in training, experience, and expertise, a data capture tool is needed that can be used reliably among researchers and graduate students (after basic training) with varying levels of experience. We have therefore identified and implemented recommendations to reduce random and systematic error. The Web-based version, currently in development, will be enhanced with hyperlink definitions for ensuring consistent reviews, a programming algorithm for automatically assigning the overall quality score, and a user’s manual or help function for clarifying statistical and measurement terminology.

The interrater reliability rates suggest that despite the course work in SLRs provided to the reviewers, more experience in the application of measurement theory was needed for novice raters (Reviewers 2 and 4). However, experienced raters (Reviewers 1 and 3), familiar with the application of measurement, demonstrated greater consistency in their interrater reliability rates. Despite using the blinded review for the second round of data extraction, reviewers might have introduced error by memory effects or by learning effects (i.e., reviewers were more proficient during the retest). Of all the 33 variables for which intrarater reliability was computed, the percentage of those for which reliability was in the good/excellent category varied from 76% for Reviewer 3 (most experienced) to 55% for Reviewer 2 (most novice). This variability also suggests using blinded reviews during both the first and second data extractions.

Limitations

Usefulness of the SPIDER tool has yet to be determined in other research groups, and our findings are therefore not generalizable at this early stage. Reducing the time for abstracting information (currently 45 min) from studies may improve rater burden. Further empirical testing of the constructs, domains, and items within other research teams is recommended.
Appendix.
Systematic Process for Investigating and Describing Evidence-based Research (SPIDER) Measurement Tool
Reviewer instructions are enclosed in brackets.

A. Background
1. Reviewer information
   1a. Reviewer’s name__________________
   1b. Date of review [mmddyy]___________
2. Source information
   2a. Study identifier [title]_____________________
   2b. Date of publication ___________________
   2c. Date of data collection _______________
3. Source type
   - Journal article
   - Book
   - Unpublished article
   - Thesis or dissertation
   - Other [specify]________________________
4. Sources of funding [check all that apply]
   - NIH
   - Other federal [list]
   - Other state [list]
   - Other [specify]________________________
   - Not applicable

B. Screening Questions
[continue if yes to (1 or 2) and (6 or 7) + yes to each of 3, 4, 5]
1. Does the source pertain to a quantitative research study?
   - Yes  □  No
2. Does the source pertain to a qualitative research study?
   - Yes  □  No
3. Does the source contain research results?
   - Yes  □  No
4. Does the source pertain to senior adults?
   - Yes  □  No
   Age definition of senior:________
5. Does the source pertain to U.S. studies?
   - Yes  □  No
6. Does the source pertain to unsafe driving (e.g., near crashes, minor crashes, driving citations or errors, fatalities, injuries)?
   - Yes  □  No
7. Does the source pertain to safe driving?
   - Yes  □  No

C. Study Descriptors
1a. Purpose [list all]________________________
1b. Purpose
   - Not stated  □  Very unclear
   - Somewhat unclear  □  Somewhat clear
   - Very clear
2a. Objectives [list all]________________________
2b. Objectives
   - Not stated  □  Very unclear
   - Somewhat unclear  □  Somewhat clear
   - Very clear
3a. Specific aims [list all]________________________
3b. Specific aims
   - Not stated  □  Very unclear
   - Somewhat unclear  □  Somewhat clear
   - Very clear
4a. Hypotheses [list all]________________________
4b. Hypotheses
   - Not stated  □  Very unclear
   - Somewhat unclear  □  Somewhat clear
   - Very clear
5a. Relevance [list all]________________________
5b. Relevance
   - Not stated  □  Very unclear
   - Somewhat unclear  □  Somewhat clear
   - Very clear
6a. Rationale [list all]________________________
6b. Rationale
   - Not stated  □  Very unclear
   - Somewhat unclear  □  Somewhat clear
   - Very clear
7a. Was existing database used?  □  Yes  □  No
   [if yes—checklist appears]
7b. Was a new database developed?  □  Yes  □  No
### D. Sampling Considerations

1. Report of sampling method (QI):
   - Reported in this source
   - Reported in another source
   - Not reported

2. Sampling methods [check all that apply]
   - Random
   - Convenience
   - Other [specify]
   - ____________________
   - ____________________
   - ____________________

3. [Qualitative studies only] Why were the subjects selected? (Rationale for selection)
   ____________________

4. Informed consent obtained? (QI)
   - Yes
   - No
   - Not discussed

5. Inclusion/exclusion criteria [if applicable]
   5a. Inclusion criteria [list all]
   5b. Exclusion criteria [list all]

### E. Rater Reliability

1a. Was intrarater reliability determined? (QI)
   - Yes
   - No
   - Not discussed

1b. Type of statistic used: ____________________
    Value of statistic: ____________________

2a. Was interrater reliability established?
   - Yes
   - No
   - Not discussed

2b. Type of statistic used: ____________________
    Value of statistic: ____________________

### F. Design Considerations

1. Sample size justification given? (QI)
   - Yes
   - No
   - Not applicable

2. Study design [choose all terms used to describe the study design]
   - Quantitative (QI)
   - Qualitative
   - Case series, case report, cross-sectional
   - Descriptive
   - Inferential/analytical
   - Randomized control trial
   - Factorial design
   - Crossover design
   - Group randomized controlled trial
   - Uncontrolled trial
   - Quasi experimental
   - Ecologic study
   - Case-control study
   - Nonrandomized control trial
   - Prospective cohort study
   - Nested case-control study
   - Retrospective cohort study
   - Concurrent retrospective and prospective cohort study
   - Time series with multiple measurements
   - Case-only study with retrospective exposure measurement
   - Case-only study with prospective exposure measurement
   - Single-subject design
   - Other [specify]

3a. Was the measurement schedule reported?
   - Yes
   - No

3b. Number of time points: __________

3c. Length of time between measurements [report for all time points]: __________

3d. Were the measurement schedules the same for all groups?
   - Yes
   - No
### G. Statistical Analysis [Quantitative Only]

1. Methods checklist [check all methods reported]
   - ANCOVA
   - ANOVA
   - Bonferroni
   - Chi square
   - Cronbach’s alpha
   - Cox proportional hazard methods
   - Discriminant analysis
   - Estimation–confidence intervals
   - Factor analysis
   - Friedman two-way ANOVA
   - ICC (intraclass correlation)
   - Kappa
   - Kruskal-Wallis
   - Logistic regression
   - MANCOVA
   - MANOVA
   - Mann-Whitney U
   - Multiple logistic regression
   - Multiple regression
   - Multivariate
   - Nonparametric tests
   - Partial correlation
   - Pearson
   - Poisson regression
   - Regression
   - Spearman
   - Structural equation models
   - Summary statistics
   - t test
   - Wilcoxon signed rank test
   - Z test
   - Other [specify] __________________________

2. Missing data
   2a. Was there a discussion about item nonresponse? (QI)  
      - Yes  No
   2b. Was there a discussion about loss to follow-up? (QI)  
      - Yes  No
   2c. How many subjects were lost to follow-up? ______
   2d. Were data imputed?  
      - Yes  No  Not discussed

3. Outliers (QI): Was detection of outliers discussed?  
   - Yes  No

4. Diagnostics (QI):  
   4a. Were model diagnostics discussed?  
      - Yes  No
   4b. Was there a sensitivity analysis?  
      - Yes  No

5. Model fit (QI): Was model fit discussed?  
   - Yes  No

6. Group equivalence (QI): Was matching or equivalence discussed?  
   - Yes  No

### H. Data Collection [Qualitative Only]

1. Was setting described?  
   - Yes  No

2. What data collection methods were used?  
   - Structured interview
   - Unstructured interview
   - Focus group
   - Observation
   - Other, specify___________________________

3. Rationale given for data collection methods?  
   - Yes  No  Not discussed

4. Data format: [check all that apply]  
   - Transcript
   - Tape recording
   - Field notes
   - Diary or journal
   - Video
   - Photographs
   - Other, specify___________________________

5. Was length of data collection session(s) given?  
   - Yes  No  Not discussed

6. Were methods altered during study?  
   - Yes  No  Not discussed
   - If so, explain why_________________________

7. Was saturation (data repetition) reached?  
   - Yes  No  Not discussed

### I. Rigor of Analysis [Qualitative Only]

1. Was there a description of the process for data analysis?  
   - Yes  No

2. Data analysis methods [check all that apply]  
   - Open coding
   - Selected coding
   - Category identification
   - Theme identification
   - Merging of themes
   - Merging of categories
   - Other, specify___________________________
3. Was original research question answered?
   - Yes  
   - No  
   Comments______________________________________________________________

4. Are the findings supported by data (e.g., verbatim responses of participants, survey results)?
   - Yes  
   - No  
   Comments______________________________________________________________

5. Are conflicting data discussed?
   - Yes  
   - No  
   Comments______________________________________________________________

6. Did the researcher(s) discuss the impact of their role(s) on data analysis?
   - Yes  
   - No  
   Comments______________________________________________________________

7. Did the use of qualitative methods lead to obtaining the stakeholders’ views and expression of the themes and concepts they stressed?
   - Yes  
   - No  
   - Not discussed  
   Comments______________________________________________________________

8. What validation methods were used? [check all that apply]
   - Triangulation
   - Member checking
   - Use of multiple analysts
   - Expert opinion
   - Theory
   - Existing literature
   - Other, specify__________________________
   - Not discussed

J. Description of Sample
1. Population__________________________________________  
   - NA  
   - Unknown
2. No. participants________________________________________
3. No. excluded__________________________________________  
   - NA  
   - Unknown
4. No. included in this analysis________________________________________
5. No. lost to follow-up________________________________________  
   - NA  
   - Unknown
6. Sample divided into groups?  
   - Yes  
   - No  
   Number of groups_____  
7. Information on group assignment
   a. Name of group
   b. Definition of group
   c. No. in group

8a. Were incentives given?  
   - Yes  
   - No  
   - Not stated
8b. Were incentives given to all groups?  
   - Yes  
   - No  
   - Not stated
9. Description given by:  
   - Entire sample
   - Group description
   - Both
10. Sociodemographic variables by whole sample or subgroups

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type of Variable</th>
<th>No. of Categories (if categorical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10a. Age</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
<tr>
<td>10b. Gender</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
<tr>
<td>10c. Race/ethnicity</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
<tr>
<td>10d. Years education</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
<tr>
<td>10e. Geographic origin</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
<tr>
<td>10f. Income</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
<tr>
<td>10g. Living status</td>
<td>Numerical</td>
<td>Categorical</td>
</tr>
</tbody>
</table>
The American Journal of Occupational Therapy

11. Numerical data [for 10a–10g]

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Mean</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Range</th>
<th>SD</th>
</tr>
</thead>
</table>

12. Category variable name / data [for 10a–10g]

<table>
<thead>
<tr>
<th>N</th>
<th>%</th>
</tr>
</thead>
</table>

13. In addition to sociodemographic variables in #10. Give number of other variables ____________

<table>
<thead>
<tr>
<th>13a. Variable description or name</th>
<th>13b. How measured</th>
<th>13c. Data set</th>
<th>13d. Instrument type</th>
<th>13e. Variable type (check all that apply)</th>
<th>13f. Validity discussed</th>
<th>13g. Reliability discussed</th>
<th>13h. Rationale discussed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>☐ Standardized</td>
<td>☐ Yes</td>
<td>☐ Yes</td>
<td>☐ Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>☐ Nonstandardized</td>
<td>☐ No</td>
<td>☐ No</td>
<td>☐ No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>☐ Unknown, not sure</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

14. Description of sample: ☐ Minimal  ☐ Moderate  ☐ Well described

K. Results

1a. Main results [list all] __________________________

1b. Secondary results available by subpopulation? ☐ Yes ☐ No

1c. Results ☐ Not clear  ☐ Somewhat clear  ☐ Very clear

L. Technical Details of Reported Results

<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Graphs or figures used?</td>
<td>2a. Graphs or figures used?</td>
</tr>
<tr>
<td>☐ Yes ☐ No</td>
<td>☐ Yes ☐ No</td>
</tr>
<tr>
<td>1b. Results given in tables?</td>
<td>2b. Results given in tables?</td>
</tr>
<tr>
<td>☐ Yes ☐ No</td>
<td>☐ Yes ☐ No</td>
</tr>
<tr>
<td>1c. Results presentation [check all that apply]</td>
<td>2c. Conceptual model presented?</td>
</tr>
<tr>
<td>☐ Means</td>
<td>☐ Yes ☐ No</td>
</tr>
<tr>
<td>☐ Percentages</td>
<td>2d. Narrative summary of thematic results provided?</td>
</tr>
<tr>
<td>☐ Correlations</td>
<td>☐ Yes ☐ No</td>
</tr>
<tr>
<td>☐ Odds ratios</td>
<td>2e. Respondents’ quotes used?</td>
</tr>
<tr>
<td>☐ Regression coefficients</td>
<td>☐ Yes ☐ No</td>
</tr>
<tr>
<td>☐ Relative risk</td>
<td>2f. Other [specify] __________________________</td>
</tr>
<tr>
<td>☐ Confidence intervals</td>
<td></td>
</tr>
<tr>
<td>☐ Credible intervals</td>
<td></td>
</tr>
<tr>
<td>☐ p values</td>
<td></td>
</tr>
<tr>
<td>☐ Medians</td>
<td></td>
</tr>
<tr>
<td>☐ Standard errors</td>
<td></td>
</tr>
<tr>
<td>☐ Standard deviation</td>
<td></td>
</tr>
<tr>
<td>☐ Variances</td>
<td></td>
</tr>
<tr>
<td>☐ Other [specify] __________________________</td>
<td></td>
</tr>
<tr>
<td>1d. Were effect sizes reported?</td>
<td>☐ Yes ☐ No</td>
</tr>
</tbody>
</table>

3. Results format: ☐ Not effective  ☐ Somewhat effective  ☐ Very effective
M. Conclusions
1a. Was there a statement of conclusions given? □ Yes □ No
1b. Main conclusions [list all] ____________________
1c. Conclusions: □ Not clear □ Somewhat clear □ Very clear

N. Limitations Cited by Authors
1. Were study limitations discussed? □ Yes □ No
2. Quantitative study limitations (QI)
   2a. Bias □ Yes □ No □ Not clearly stated
   2b. Confounding □ Yes □ No □ Not clearly stated
   2c. Internal validity □ Yes □ No □ Not clearly stated
   2d. External validity □ Yes □ No □ Not clearly stated
   2e. Power □ Yes □ No □ Not clearly stated
   2f. Experiment wide error □ Yes □ No □ Not clearly stated
   2g. Other [specify] ____________________________

3. Qualitative study limitations
   3a. Did the researcher(s) discuss the impact of their role(s) on study design? □ Yes □ No Comments________________
   3b. Did the researcher(s) discuss the impact of their role(s) on study recruitment? □ Yes □ No Comments________________
   3c. Did the researcher(s) discuss the impact of their role(s) on data collection? □ Yes □ No Comments________________
   3d. Did the researcher(s) discuss ethical issues raised by the study? □ Yes □ No
   3e. Ethical issues discussed [check all that apply] □ Privacy □ Confidentiality
   □ Security □ Institutional Review Board
   3f. Was credibility discussed (participants’ judgment of believability of results)? □ Yes □ No
   3g. Was transferability discussed (application of the research to other groups)? □ Yes □ No
   3h. Was dependability discussed (the influence of setting and context on the results)? □ Yes □ No
   3i. Was confirmability discussed (how were the results checked)? □ Yes □ No

O. General Impressions of Study
1. Domains discussed (check all that apply):
   □ Social □ Primary focus □ Secondary focus □ Not discussed
   □ Epidemiological □ Primary focus □ Secondary focus □ Not discussed
   □ Health □ Primary focus □ Secondary focus □ Not discussed
   □ Behavioral □ Primary focus □ Secondary focus □ Not discussed
   □ Ecological □ Primary focus □ Secondary focus □ Not discussed
   □ Other [list _____________________________________]
   □ Primary focus □ Secondary focus □ Not discussed

P. Implications
1a. Implications [list all] ________________________
1b. Type of implication (check all that apply)
   □ Economic □ Policy □ Practice
   □ Social □ Psychological □ Research
   □ Other (specify) _________________________

Q. Quality Score [Assigned by Reviewer]
1. Overall rating of source quality:
   0 1 2 3 4 5 6 7 8 9 10
   Very poor Excellent

R. Keywords
1a. Keywords identified in the article
   □ Epidemiology □ Health □ Medications
   □ Main medical conditions □ Physical health □ Mental health
   □ Comorbidities □ Stake-holders □ Client perspective
   □ Existing datasets □ Retrospective study □ Prospective study
   □ Behavior □ Environment/ ecology □ Prevention
   □ Risk factors □ Cessation □ Morbidity
   □ Fatality □ Impairment □ Evaluation
   □ Assessment □ Outcome □ High risk
   □ Low risk □ Mortality □ Other
1b. Keywords identified by the reviewer
   (same checklist shown in R.1a.)