How valid and reliable are patient satisfaction data? An analysis of 195 studies

JOHN SITZIA

Research and Development Department, Worthing and Southlands Hospitals NHS Trust, Worthing, UK

Abstract

Objective. To assess the properties of validity and reliability of instruments used to assess satisfaction in a broad sample of health service user satisfaction studies, and to assess the level of awareness of these issues among study authors.

Design. Examination and analysis of 195 papers published in 1994 in 139 journals. The following databases were searched: British Nursing Index, CINAHL, EMBASE, MedLine, Popline, and PsycLIT.

Main measures. Number and types of strategies used for content, criterion, and construct validity, and for stability and internal consistency. Associations between validity/reliability and other study characteristics.

Results. Eighty-nine (46%) of the 195 studies reported some validity or reliability data; 76 reported some element of content validity; 14 reported criterion validity, with patient's intent to return the most commonly used criterion; four reported construct validity. Thirty-four studies reported internal consistency reliability, 31 of which used Cronbach’s coefficient $\alpha$; eight studies reported test–retest reliability. Only 11 studies (6% of the 181 quantitative studies) reported content validity and criterion or construct validity and reliability. ‘New’ instruments designed specifically for the reported study demonstrated significantly less evidence for reliability/validity than did ‘old’ instruments.

Conclusion. With few exceptions, the study instruments in this sample demonstrated little evidence of reliability or validity. Moreover, study authors exhibited a poor understanding of the importance of these properties in the assessment of satisfaction. Researchers must be aware that this is poor research practice, and that lack of a reliable and valid assessment instrument casts doubt on the credibility of satisfaction findings.

Keywords: data collection, meta-analysis, patient satisfaction, psychometrics, questionnaires, reproducibility of results
and must use appropriate formats to measure those factors, this is referred to as content validity. Second, the test must demonstrate the extent to which it produces results which reflect true values. Traditionally there have been two approaches to this.

Criterion validity is the correlation of the new scale with some other measure of the trait under study, ideally an existing validated ‘gold standard’ which is well accepted in the field. If the two measures are administered at the same time, this is referred to as concurrent validity. If the criterion is not available until some time in the future, this is called predictive validity. This type of validity is important, for example, in screening measures and prognostic measures.

When no valid criterion exists, construct validity may be tested. Construct validity refers to the extent to which the results obtained using a measure concur with the results from the underlying theoretical construct. Construct validity is complex, but essentially consists of a three-step process:

(i) first, using research evidence as far as possible, a hypothetical relationship is built between our construct and an observable. For example, a relationship between satisfaction with waiting time at a clinic (our construct) and ‘walk-outs’ from the clinic (the observable) could reasonably be predicted;

(ii) second, those items that relate to the construct by statistical analysis are identified. For example, we might find a high correlation between the items measuring satisfaction with waiting time at reception, waiting time for doctor, and waiting environment;

(iii) the third step is to perform studies to determine the extent to which the supposed measures of the concept produce the predicted results. A number of tests are may be used for this. The two most common are:

- discrimination or ‘extreme groups’. Typically this test identifies two groups who in theory would be likely to produce strongly differing results in terms of the observable. Validity would be supported if the results did indeed differ in the predicted direction;

- convergent/discriminant method. In this method validity is determined by the extent to which our scale – the observable) could reasonably be predicted;

Previous work suggests that most satisfaction studies have weak reliability and validity [8,9]. Content validity is seldom established, beyond ‘consensus of experts’ or judgement of face validity, and typically neither criterion nor construct validity are considered. One review reported that reliability was measured in around 40% of instruments [9].

Although useful, these reviews have not provided a systematic assessment of validity and reliability in a broad sample of health service user satisfaction studies. This was the primary aim of the current study. A secondary aim was to assess the level of awareness of these issues among study authors.

Methods

Selection of papers

Satisfaction study reports have appeared in the health literature for at least 40 years. In recent years the annual output of papers indexed as satisfaction-related has reached several hundred. So, as a comprehensive review would involve scrutiny of thousands of papers, this study aimed to use a representative sample.

The sample consisted of all satisfaction reports published in one year, written in the English language, and indexed in the following databases: British Nursing Index, CINAHL, EmBase, MedLine, Popline, and PsycLIT. The following subject headings, and combinations thereof, were used: patient satisfaction; consumer satisfaction; patient acceptance of health care; health services research; quality of health care; quality assurance, health care; delivery of health care; health facilities; data collection; health surveys; health care surveys; questionnaires; interviews; inpatients; outpatients.

1994 was chosen as the subject year, as the number of published papers appeared to peak in this year [3]. No attempt
was made to sample the grey literature of unpublished reports, theses, and local surveys. Given that local surveys are associated with low methodological rigour [10], and adopting the assumption that studies with poor methodology are rejected for publication, this strategy therefore attempted to identify a sample representing ‘good quality’ reports.

The sole criterion for inclusion in this analysis was that the published report included the results of an explicit assessment of health service user satisfaction; explicit means that the investigators stated that satisfaction was a dependent variable under study. Three types of papers identified by the search strategy were excluded:

(i) papers which did not report an assessment of user satisfaction: editorials, letters, discussion papers, comments, critiques, review articles, non-patient assessments of satisfaction, and non-satisfaction assessments of care quality;
(ii) studies in which the sole or primary aim was the development of a satisfaction assessment instrument. These studies were excluded as the aim of the current study was to examine instruments actually in use, rather than in development;
(iii) studies which could reasonably be interpreted as satisfaction-related, but in which satisfaction was not explicitly identified as the construct under study, such as investigations of patients’ ‘concerns’, ‘feelings’, ‘perspectives’, ‘perceptions’, ‘experiences’, ‘reactions’, ‘attitudes’, ‘views’, ‘thoughts’ or ‘opinions’.

Data analysis

Data were extracted and analysed in these stages.

Frequency of these study characteristics

- Study provenance (country);
- clinical context (i.e. the service under assessment);
- primary profession of investigators.

Frequency of these instrument characteristics

- Data collection procedure;
- type of data produced (qualitative or quantitative);
- origin of the instrument. Three origin groups were identified: new – an instrument developed for the reported study; old-unmodified – an existing instrument used with no modification; old-modified – an existing instrument modified for the study.

Instrument development and testing data

Instrument development and testing data were extracted from the study report – the primary report – and from a maximum of two secondary sources cited in the primary report (e.g. a full report, an instrument development report).

For study instruments producing principally quantitative data, data relevant to the following elements were extracted.

Content validity

- Strategies used for item generation;
- whether or not the instrument used a discriminatory scale, defined in this case as a scale with at least five response points. Five is the minimum number of categories suggested by an analysis of the relationship between number of categories and reliability, described by Streiner and Norman [6];
- inclusion of open questions for comments [11,12];
- number of items;
- strategies used for content testing.

Criterion validity

- Criteria used;
- strength of correlation between criteria and satisfaction ratings.

Construct validity

- Hypothesis used;
- procedure used to identify relevant items;
- analysis used to test the hypothesis.

Internal consistency

- Test used;
- test result.

Stability

- Test used;
- test result.

For study instruments producing principally qualitative data, any relevant data was extracted and frequency of data reporting calculated.

Minimum validity and reliability

Quantitative studies exhibiting a minimum level of validity and reliability were identified. The minimum for content validity was defined as evidence of at least one strategy each for item generation and content testing. The minimum for criterion or construct validity and for reliability was simply evidence of some work on that property. Studies which provided evidence for all three properties were defined as meeting the overall minimum level of reliability and validity.

Use of the $\chi^2$ test

For quantitative studies, the $\chi^2$ test was used to examine associations between reliability/validity data and those study/instrument characteristics which might reasonably be expected to have a bearing on the quality of the study instrument. The characteristics were:

- whether or not the study team included an academic (academic teams; on the assumption that academic authors, generally with a higher level of research training and experience than non-academic authors, have greater knowledge of reliability/validity issues than non-academics);
Awareness of reliability and validity

It was noted whether or not the text displayed any awareness on the part of the authors of reliability and validity issues. Awareness was defined as any mention of these issues, typically in the Discussion section of the report, and typically in terms of (i) the limitations of the study with respect to reliability and validity, and/or (ii) as issues deserving further work.

Results

Study characteristics

The search identified a total of 198 eligible papers. Three studies were found to be duplicate reports (i.e. identical results reported in two journals); one version of each was excluded, leaving 195 papers. When papers reported results from more than one survey, data only from the first-reported survey were used. Reports were published in a total of 139 journal titles. A full list of reviewed papers is available from the author. The list is divided into two groups: papers providing no instrument reliability or validity data (\(n=106\)) and papers providing some instrument reliability or validity data (\(n=89\)).

Descriptive data for study characteristics are presented in Table 1. Studies examining ‘surgical or medical procedure’ typically assessed satisfaction with a specific intervention, whereas studies in all other categories typically assessed satisfaction with a service or unit.

Instrument characteristics

Satisfaction data were collected by self-report questionnaire in 125 studies (64%), by structured interview or face-to-face questions in 55 studies (28%), and by unstructured or semi-structured interview in nine studies (5%). Six reports (3%) did not state the method of data collection.

Satisfaction data were primarily quantitative in 181 studies (93%), and primarily qualitative in 14 studies (7%).

Of the sample, 81% (158 studies) assessed satisfaction using a new instrument; 17 studies (9%) used an old-unmodified instrument; and 20 studies (10%) used an old-modified instrument.

Instrument development and testing

In this and subsequent sections, ‘data’ means data relevant to the reliability and/or validity of the satisfaction assessment instrument.

Twenty-four primary papers referenced one secondary source, and two referenced two secondary sources. Of these 28 secondary sources, four were inaccessible (unpublished, or unpublished in English), six reported no relevant data, and 18 (64% of secondary sources, secondary to 17 primary papers) reported some relevant data.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>(n)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>83</td>
<td>43</td>
</tr>
<tr>
<td>UK</td>
<td>53</td>
<td>27</td>
</tr>
<tr>
<td>Other Europe</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>Canada</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Australia or New Zealand</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Clinical context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surgical or medical procedure</td>
<td>47</td>
<td>24</td>
</tr>
<tr>
<td>Hospital inpatient</td>
<td>41</td>
<td>21</td>
</tr>
<tr>
<td>Primary care</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>Hospital outpatient/casualty</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>Mental health services</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Maternity care</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>Home care/palliative care</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Dental care</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Profession of investigating team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>68</td>
<td>35</td>
</tr>
<tr>
<td>Academic</td>
<td>53</td>
<td>27</td>
</tr>
<tr>
<td>Mixed academic/clinical</td>
<td>66</td>
<td>34</td>
</tr>
<tr>
<td>Other (charity, consumers’ association)</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>No data</td>
<td>1</td>
<td>&lt;1</td>
</tr>
</tbody>
</table>

Some data were available for 89 studies (46%) (Table 2). For the old-modified group, data relevant to the psychometric properties of the original instrument were available in either the primary or a secondary source in nine cases (45% of the group).

Of the 89 studies with data, 82 were primarily quantitative and seven were primarily qualitative. Some data were available for 45% (82/181) of quantitative studies and 50% (7/14) of qualitative studies (\(\chi^2 = 0.1\), not significant).

Data from quantitative studies

Summary data are presented in Table 3.

Content validity

Some element of content validity was reported by 76 studies. Thirty-seven (49% of this group) reported at least one strategy for each of item generation and content testing.

(i) Item generation. Sixty studies noted item generation; 32 (53%) used only one source (e.g. patients), 16 (27%) used two, 10 (17%) used three, and two (3%) used five.

(ii) Rating scale. Ten (5%) of the 181 quantitative studies provided no indication of the scale used to assess satisfaction. Of the other 171, 32 (19%) used principally nominal scale items, 134 (78%) used principally ordinal scale items, and five (3%) used principally interval scale items. One hundred and sixteen papers...
Table 2 Sources of validity and reliability data for three groups of satisfaction assessment instruments. Unless stated otherwise, cells show number (n) and percentage (%) within each instrument group

<table>
<thead>
<tr>
<th>Instrument group</th>
<th>Data source</th>
<th>Primary only</th>
<th>Secondary only</th>
<th>Both primary and secondary</th>
<th>None</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>New</td>
<td></td>
<td>61</td>
<td>39</td>
<td>0</td>
<td>95</td>
<td>158</td>
</tr>
<tr>
<td>Old-unmodified</td>
<td></td>
<td>2</td>
<td>12</td>
<td>5</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>Old-modified</td>
<td></td>
<td>9</td>
<td>45</td>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>72</td>
<td>37</td>
<td>11</td>
<td>6</td>
<td>195</td>
</tr>
</tbody>
</table>

Table 3 Number and percentages of study reports providing data for each type of validity and reliability

<table>
<thead>
<tr>
<th>Property</th>
<th>% Quantitative papers with some reliability data (n = 82)</th>
<th>% All quantitative or validity data papers (n = 181)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content validity</td>
<td>76</td>
<td>42</td>
</tr>
<tr>
<td>Item generation</td>
<td>60</td>
<td>33</td>
</tr>
<tr>
<td>Health professionals</td>
<td>32</td>
<td>18</td>
</tr>
<tr>
<td>Literature</td>
<td>37</td>
<td>20</td>
</tr>
<tr>
<td>Patients/users</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Managers</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Relatives/carers</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Ethicist</td>
<td>1</td>
<td>&gt;1</td>
</tr>
<tr>
<td>Complaints</td>
<td>1</td>
<td>&gt;1</td>
</tr>
<tr>
<td>Content testing</td>
<td>53</td>
<td>29</td>
</tr>
<tr>
<td>Discriminatory power</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Components</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Face validity</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Pilot study</td>
<td>41</td>
<td>23</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Construct validity</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>34</td>
<td>19</td>
</tr>
<tr>
<td>Stability</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Test–retest</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Six papers reported that the instrument had been examined for face validity; five used health professionals, one used patients.

Criterion validity

Criterion validity was measured in 14 studies.

(i) Criteria used. The most common approach was correlation of a satisfaction index (typically the total score from all satisfaction items) with a single question on the patient’s intent to return (Table 4).

(ii) Strength of correlation. See Table 4.
Table 4 Criteria used for tests of criterion validity of the satisfaction assessment instrument, showing number of studies which adopted the criteria (n) and the median correlation (r) between instrument and criteria. More than one criterion could be used.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>n</th>
<th>Median (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient satisfaction</td>
<td>17</td>
<td>0.86</td>
</tr>
<tr>
<td>(different instrument)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual items in same instrument</td>
<td>2</td>
<td>0.72</td>
</tr>
<tr>
<td>Global quality (one item)</td>
<td>2</td>
<td>0.68</td>
</tr>
<tr>
<td>Recommend to others</td>
<td>3</td>
<td>0.56</td>
</tr>
<tr>
<td>Health professionals’ rating of patient’s satisfaction</td>
<td>4</td>
<td>0.51</td>
</tr>
<tr>
<td>Intent to return</td>
<td>5</td>
<td>0.38</td>
</tr>
<tr>
<td>Functional status</td>
<td>1</td>
<td>0.29</td>
</tr>
<tr>
<td>Health-related quality of life</td>
<td>2</td>
<td>0.26</td>
</tr>
<tr>
<td>Psychological status</td>
<td>4</td>
<td>0.26</td>
</tr>
<tr>
<td>Number of problems reported</td>
<td>1</td>
<td>no data</td>
</tr>
</tbody>
</table>

Construct validity

Data were available for four studies (4% of studies with data, 2% of quantitative studies). None of these provided data for each aspect of construct validity testing. Two studies used the same instrument, the La Monica–Oberst Patient Satisfaction Scale (LOPSS). Both cited a secondary paper; however, this reported a factor analysis to examine the robustness of the three dimensions of satisfaction included in the LOPSS, i.e. testing of components for content validity [13]. The third study provided details of the properties predicted to determine satisfaction but provided no details of the tests used to examine the hypothesis and no details of results. The remaining study assessed satisfaction via the health care orientation domain of the Patient Adjustment to Illness Scale (PAIS). Discriminant ability of the PAIS was tested by comparing results from healthy and ill people, but the groups showed no significant difference in health care orientation (i.e. validity was not supported)[14].

Internal consistency

Internal consistency data were available for 34 studies (38% of studies with data, 19% of quantitative studies).

(i) Test used. Estimation using Cronbach’s α was the most popular approach (Table 5), used in a total of 31 studies (one study calculated both subscale α and total scale α).

(ii) Test result. A coefficient a value of 0.80 or more was recorded in 26 cases (84% of the 31 studies) (Table 5).

Stability

Eight studies (9% of studies with data, 4% of quantitative studies) reported test–retest reliability. Two further studies reported inter-rater reliability.

(i) Test used. Six studies used correlation, and three measured agreement using the κ statistic; the remaining study did not provide data.

(ii) Test result. Studies using correlation (n=6): median r=0.71, range = 0.67–0.86. Studies using κ (n=3): one reported a mean κ of 0.51, one reported κ as ‘more than 60% for most questions’, and one reported that agreement ‘ranged from 62% to 84%’.

Data from qualitative studies

Seven qualitative studies provided some data (50%). All data were concerned with the content validity of the instrument (more than one strategy could be used); two studies consulted health professionals in the process of instrument development, four referred to previous research findings, one consulted patients, and two conducted a pilot study.

Minimum level of evidence for reliability and validity

The minimum level of evidence was provided for content validity by 37 studies, for criterion or construct validity by 16 studies, and for reliability by 37 studies.

Eleven studies (6% of the 181 quantitative studies) met the criterion for an overall minimum. Of these:

- three studies used a new instrument [15–17];
- One study [29] used a modified version of an existing instrument [30].
Relationships between study/instrument characteristics and reliability/validity data

The proportion of studies with data was 55% among papers generated by academic or mixed academic/clinical authors, as opposed to 28% among papers generated by clinical authors ($\chi^2 = 13.5, P < 0.01$). This pattern was also evident for the minimum content validity (23% versus 11%, $\chi^2 = 3.9, P < 0.05$), and for minimum reliability (25% versus 7%, $\chi^2 = 9.8, P < 0.01$). No significant difference in proportions was found for minimum construct/criterion validity (9% and 6%) or the overall minimum (7% and 1%).

Of studies which used an old-unmodified instrument 94% provided some data compared with 40% of studies which used either a new or old-modified instrument ($\chi^2 = 18.0, P < 0.01$). This pattern was also evident for minimum content validity (59% versus 15%, $\chi^2 = 19.2, P < 0.01$), for minimum construct/criterion validity (59% versus 3%, $\chi^2 = 63.3, P < 0.01$), for minimum reliability (88% versus 12%, $\chi^2 = 58.1, P < 0.01$), and for the overall minimum (41% versus 2%, $\chi^2 = 44.2, P < 0.01$).

Country of origin: 69% of papers emanating from Canada provided some data, compared with 43% for the UK, 48% for the USA, 40% for other Europe, 50% for Australia/New Zealand and 11% for other countries. Of the papers from Canada 15% met the overall minimum criterion, compared with 6% from the UK and USA, 4% from other Europe, and 0% from other sites.

Clinical context: 67% of papers from maternity care and 63% from mental health provided some data, compared with 13% for surgery and medical procedures.

Awareness of reliability and validity issues

Of the 106 studies presenting no reliability or validity data, three (3%) noted the need for instrument testing. Of the 89 studies presenting some data, nine (10%) demonstrated an awareness (as defined in the Methods section above).

Discussion

This study aimed to examine the validity and reliability of satisfaction data through an investigation of assessment instruments used in a representative sample of published studies. The search strategy was of key importance, and could be criticized for (i) searching only a small number of databases, and (ii) selecting only English-language reports. However, given the eventual size of the sample and the range of journals included, I believe that it is unlikely that the strategy has limited the generalizability of the results. Another consideration is the exclusion of papers which did not explicitly assess satisfaction. Many of these examined constructs which both logic and research evidence would associate with satisfaction; however, as this study investigated satisfaction assessment instruments, it was important that each report stated explicitly that this was indeed the construct under assessment, whether or not the authors displayed any appreciation of the complexities of the construct.

The finding that most studies collected data via a self-report questionnaire is in line with anecdotal evidence. However, the very low proportion of studies adopting a primarily qualitative approach was unexpected and also disappointing; qualitative research aims to develop theory inductively and patient satisfaction research lacks a good theoretical base. These data support the proposition that most satisfaction research is evaluative rather than descriptive, the primary aim being to judge service quality [31]. However, it is recognized that for evaluative data to be meaningful they must be considered against patients’ expectations [2,32–34]; as van Maanen [35] wrote: “‘quality’ is an abstraction defining the margin between desirability and reality”. Expectations are rarely assessed in relation to satisfaction data [36,37], despite a strong argument that this should happen [2].

The finding that 80% of studies produced a new satisfaction assessment instrument, and that a further 10% modified an existing instrument, was not unexpected. Most studies are context-specific, and authors may feel that existing instruments include irrelevant items or exclude relevant items or that the mix of items and dimensions is unsuitable. This is also supported by the fact that only 13 unmodified instruments were used; four studies in nursing care used two different versions of Risser’s Patient Satisfaction Instrument [13,25], and three studies in mental health used the eight-item version Client Satisfaction Questionnaire [21,38]. It is clearly difficult to ascertain the degree to which researchers are justified in deciding that an existing instrument is unsuitable for their particular study; however, prospective researchers would be advised (i) to make explicit the reasons for their choice of instrument, and (ii) to ensure that the validity and reliability of any new or modified instrument is equal to that of existing, comparable instruments.

An important finding was that new or modified instruments were less likely than unmodified instruments to display validity, reliability or an overall minimum. It is worth repeating that 60% of studies using a new instrument reported no reliability or validity data whatsoever; clearly this is unacceptable research practice. Academic teams may take little solace from the fact that they were significantly more likely than clinical teams to report some aspects of instrument testing, given that only 6% of academic studies reached the overall minimum.

It is generally argued that modification of an existing instrument will affect the validity and/or reliability of that instrument. However, no guidance exists on the degree of modification necessary for re-evaluation to be required; it would not seem sensible that a change in the wording of a single item [e.g. 25] be equated with the reduction of a 50-item questionnaire to 16 items [e.g. 39] or with the substitution of a yes/no scale for the original 5-point scale [e.g. 40]. To complicate matters further, purists would agree with Nunnally and Bernstein [5] that one validates the use to which a measuring instrument is put rather than the instrument itself. In this case, even well-tested instruments should be re-evaluated if the subject group or study context is different.
to that in which the instrument was tested, but this position hardly encourages the use of reliable and valid instruments, quite apart from lacking demarcation. In conclusion, two suggestions might be proposed: (i) whenever possible an existing reliable and valid instrument is used without modification; (ii) where modification is necessary, researchers should make an explicit assessment of the modification to determine whether or not further testing is needed.

Many definitions of validity can be found in the health sciences literature. Most suggest that validity means accuracy. In objective measurement, the accuracy of an instrument is the closeness of the measurements to their true values. However, application of this definition in the assessment of satisfaction is far from straightforward. A foremost problem is that the construct ‘patient satisfaction’ is ill-defined (largely because effort has focused on the production of satisfaction data at the expense of conceptual development) [3,34]. As a consequence, it is difficult to judge content validity when we do not know which are all the relevant content areas, it is difficult to judge criterion validity when we have no idea if the criterion used is in fact congruous with satisfaction, and it is difficult to judge construct validity when only minimal evidence suggests whether or not two groups would be expected to express different levels of satisfaction with a particular variable. For this reason the primary aim of this study was to report data on the strategies used to reach validity, with only secondary comment on the quality of those strategies.

No objective procedures to examine content validity appear in the literature. This study used a bilateral approach, with examination of (i) item generation, and (ii) item testing. Item generation was the most commonly-documented strategy in this study, being noted in one-third of the sample. Reference to previous literature must be regarded as a fundamental strategy, and therefore it was surprising that 80% of quantitative studies made no mention of this strategy. It is equally disappointing that service users were involved in item generation in only 11% of quantitative studies, and that carers or relatives were involved in only 3%. Furthermore, a small number of researchers reported the assessment of ‘face validity’, despite this being a weak and unscientific approach [41]. Researchers who rely too heavily on the opinions of health workers in the endeavour towards content validity should note the view that health service users and health workers/managers can have very different perceptions of which aspects are important to quality of care [42]. It is also important to note that many studies used only one or two strategies, despite the fact that many commentators have advocated a multi-lateral approach to item generation, with a particular emphasis on the desirability of qualitative, exploratory strategies at this stage [43,44].

A pilot stage is useful in assessing validity [45], and it was encouraging to find that this was a common strategy. It was also encouraging that some authors examined the discriminatory power of the scale, as it is important that item formats are appropriate to the trait under examination. The 5-point adjectival scale format is plainly the most popular scale in current use, appearing in at least one-third of quantitative reports in this sample.

It is interesting to note that of the 76 studies which reported some strategy towards content validity, only 14 – less than one in five – explicitly recognized the aim as content validity. Content validity is a fundamentally important concept in instrument development, and it is important that researchers demonstrate an understanding of the concept rather than adopting strategies by rote.

Criterion validity is strictly defined as the outcome of a comparison of results from the instrument under test with those from a validated scale measuring the same construct [6]. By this definition, only two studies in this sample measured criterion validity; the other 12 which purported to measure criterion validity simply compared satisfaction with a different indicator which they judged to be consonant with satisfaction. This is understandable given that the use of a single indicator appears elsewhere in the satisfaction literature [8]. Less excusable is the fact that citations or even an explicit argument supporting the use of the indicator were rarely provided; it was simply assumed that the indicator was valid. Informed – rather than unwitting – choice of a valid indicator is both interesting and problematic, the problem, as noted earlier, caused by the lack of definition of the construct ‘health service user satisfaction’. Rubin [8] raised many points relevant to the complex issue of whether or not satisfaction is an appropriate measure of quality of care, not least among these being that correlations are most likely to be strongest between specific aspects of both satisfaction and quality of care, rather than between the constructs as a whole. Other commentators agree that indicators such as patient progress and effects of treatment are concepts distinct from satisfaction [31], concluding that such muddling ‘points to the typical absence of conceptual validations of what constitutes patient satisfaction’ [31, p.60]. This consensus seems to be supported by the current results. Certainly the mean correlations for health-related indicators are relatively poor, as low as $r=0.26$ for psychological status.

The most popular criterion for validity testing in this sample was ‘intent to use the same facility again’, producing a moderate mean correlation ($r=0.51$). This criterion appears to be favoured by a range of commentators [8,31], and indeed is supported by evidence which suggests that behavioural intentions are directly affected by satisfaction [46]. Also popular in this sample was comparison with health professionals’ ratings of patient satisfaction, which produced a slightly higher correlation than that for intent to return. This criterion has been used previously and found to produce fairly good agreement [8], but nevertheless is rejected by others who argue that patients and professionals have different perspectives, use different evaluation criteria, and subsequently that ratings should not be expected to agree [31,42]. This is an interesting area which deserves further investigation, not least as to whether or not use of a single criterion provides criterion validity.

Only four studies in this sample used an instrument which had undergone construct validity testing, and none carried out the three-step procedure [7]. These data provided no useful information for other researchers wishing to test a satisfaction hypothesis.
The reliability of a test relies on the variance of the true scores. As the true score changes from one population to another, so reliability should be tested in each population. Some commentators regard internal consistency testing as more powerful than stability [5], others propose the opposite [6]. Calculation of internal consistency has become popular, perhaps partly because it requires no additional data collection. This method was reported in approximately 20% of studies in this sample, a proportion similar to that found by Rubin [8] in a review of 59 papers which excluded studies with poor methodologies. A value of 0.80 for coefficient \( \alpha \) is a suggested minimum for group studies [5,6], and was achieved in over 80% of cases in this sample. However, it is important to note that the calculation of an overall \( \alpha \) in multi-dimensional instruments is misguided: we would hope for a high intra-subscale correlation, but not necessarily a high inter-subscale correlation, as the dimensions of satisfaction are thought to be fairly independent.

The findings from this study support previous conclusions that the test–retest approach is used less commonly than the internal consistency approach, and that when it is, correlation is the statistic erroneously used rather than \( r \) [47].

Perhaps the most telling finding in this study is that only 6% of quantitative studies used instruments which demonstrated the overall minimum level of evidence for reliability and validity. Moreover, this must not be interpreted as meaning that reliable and valid instruments were used in this 6%; this is simply the proportion of studies which reported some data for each property. Equally depressing was the very low number of study authors who acknowledged the limitations of the instrument used; only 3% in the group of 106 studies that reported no validity or reliability data whatsoever. Broadly speaking, these data suggest that when an attempt at psychometric evaluation does take place, however half-hearted and whatever the results, the large majority of researchers appear to deem the process as itself constituting reliability or validity. It is interesting to note that in many studies satisfaction was assessed as one of a battery of indicators, and that the instruments for the other indicators were often well-tested, valid, and reliable. Furthermore, in several papers the authors explicitly acknowledged the properties of the other instruments, yet made no mention of the properties of the satisfaction instrument [e.g. 48,49]. I would argue that the reason for this inconsistency comes from a belief among health researchers that satisfaction is a ‘soft’ indicator, meaning that the concept is regarded as easy to understand and the trait easy to assess. This belief is puzzling, as there seems no good reason why satisfaction should be treated differently to other socio-psychological phenomena, such as quality of life, assessment instruments for which typically undergo more thorough development and testing. A comparison of satisfaction as assessed by valid and non-valid instrumentation would be an interesting and worthwhile experiment.

This discussion has focused exclusively on data from quantitative studies. It would be wrong to make generalizations based on the small number of qualitative papers in this sample; however, it does seem likely that qualitative researchers are equally neglectful of reliability and validity. Prospective researchers should note that qualitative approaches must also be justified in these respects [50–52].

It is important to remember that this study excluded papers in which the primary aim was instrument development. It is encouraging to note that at least 20 papers published in 1994 were excluded on this criteria, suggesting that a greater number of well-developed and tested instruments may appear in the future. Readers searching now for an acceptable measure are advised to consider first those instruments which in this review were found to demonstrate a minimum level of reliability and validity (described in Results).

It has been clearly emphasized that the components of reliability and validity determine the degree of credibility that will be, or should be, given to research findings [52]. Results from this study suggest that a large majority of published investigations of patient satisfaction use instruments which are untested or barely tested. It must be concluded that the results from such studies lack credibility.

References


42. Donabedian A. The Definition of Quality and Approaches to its Assessment. Ann Arbor, MI: Michigan Health Administration Press, 1980.


Accepted for publication 3 March 1999