Reliability Analysis in Therapeutic Research: Practice and Procedures

Kenneth J. Ottenbacher, Scott D. Tomchek

Key Word: assessment technics

Twenty studies examining the reliability of assessment devices and outcome measures in therapeutic research were reviewed and analyzed. The 20 investigations contained 215 quantitative reliability values published in either the American Journal of Occupational Therapy or Physical Therapy during the past 5 years. The reliability studies were classified as interrater, intrarater, test-retest, or internal consistency. Examination of interrater reliability accounted for 41% of all reported reliability values. Studies published in Physical Therapy were more likely to be concerned with test-retest reliability, whereas studies published in the American Journal of Occupational Therapy more often focused on interrater reliability. Examination of the data revealed that the intraclass correlation coefficient (ICC) was the most frequently reported estimate of reliability, accounting for 57% of all reported reliability coefficients. Further review of the results indicated that Pearson product-moment correlations and percentage of agreement indexes accounted for 22% of all reliability values reported in the studies examined. The Pearson product-moment correlation measures association or covariation among variables, but not agreement, and percentage agreement indexes do not correct for chance agreement. The argument is made that product-moment correlations and percentage agreement indexes are inadequate measures of interrater, intrarater or test-retest agreement. They should be used and interpreted with caution.

The importance of research in rehabilitation was recently acknowledged by the federal government through the establishment of the National Center for Medical Rehabilitation Research at the National Institutes of Health (NIH). The mandate of the center is to conduct and support research and related activities that will improve the rehabilitation services provided to consumers with a disability. In developing the research goals for the center, the NIH formed a task force that identified nine focus areas of research (NIH Task Force, 1990). One such area was the development of assessment procedures to document the effectiveness of rehabilitation interventions. Specifically, the Report of the Task Force on Medical Rehabilitation states that "a foremost priority is the thorough evaluation of existing and newly developed rehabilitation outcome measures. In particular, the meaningfulness, reliability and validity of these measures must be ascertained" (NIH Task Force, 1990, p. 156).

Although the concepts of reliability and validity are widely discussed in the professional rehabilitation literature, considerable confusion exists regarding their application and interpretation (Knapp, 1985). This situation appears particularly evident in the computation and interpretation of reliability coefficients.

Reliability has been defined as the consistency or stability of empirical indicators between raters or from one measurement to another (Waltz, Strickland, & Lenz, 1991), as the extent to which measurement is free of random errors (Martin & Bateson, 1986), and as the ratio of true score variance to observed score variance (Lord & Novick, 1968; Magnusson, 1966). Basically, reliability is the consistency of a measurement (American Psychological Association, 1985; Bartko & Carpenter, 1976; Rothstein, 1985). This last definition implies that consistency will be present when all conditions, to the best of the researcher's knowledge, are held constant (Rothstein, 1985).

Berk (1979) argued that the term reliability, as commonly used in the research literature, is a misnomer. This technical misconception of reliability is most frequently found in studies of interrater reliability in which two or more observers are making judgments regarding subject performance. Berk (1979) contended that the standard psychometric definition of reliability, based on variance components of observed and true scores, is not considered in the analysis of interrater reliability. He suggested that measures of interrater reliability are more accurately described as estimates of agreement between or among raters. He further argued that measures of covariation between raters, such as the commonly used Pearson product-moment correlation coefficient r, are mistakenly employed in these cases and erroneously interpreted as measures of direct agreement between or among raters. This problem is illustrated in a later section of this article.

Another commonly encountered problem in the ap...
plication and interpretation of reliability or agreement is
the use of special procedures that are not generalizable
and that make comparisons with other studies difficult. A
failure to account for chance agreement when reporting
reliability values and the reporting of a technique that
compares mean values or changes in proportion across
groups as an estimate of either reliability or agreement
are other examples of problems in the analysis and inter-
pretation of reliability studies.

The purpose of the present investigation was to ex-
amine and compare information for reliability research
published in the American Journal of Occupational
Therapy (AJOT) and Physical Therapy over the last 5
years. The study was designed to provide information
concerning the frequency and type of reliability proce-
dures reported in the therapeutic research literature and
to examine issues related to the interpretation of reliabil-
ity coefficients.

Method

Reliability studies were operationally defined as empiri-
cal investigations published in AJOT or Physical Therapy
reporting the results of a study classified as examining
intrarater reliability (agreement), test-retest reliability, in-
trarater reliability, or internal reliability (consistency). To
be included in this review and analysis, the study had to
identify the estimation of reliability as one of the primary
objectives or focus areas of investigation. This criterion
was generally met through the selection of studies that
contained the term reliability in the title or emphasized
the reliability analysis and results in the abstract. Oper-
tional definitions for classifying a particular reliability
assessment procedure were adapted from Bailey (1991)
and appear below.

Intrarater or Intrarater Reliability. Intrarater stud-
ies included those in which the goal was to determine the
extent to which two or more examiners agreed on the
performance or rating of a particular subject or group of
subjects. For example, Deitz, Tovar, Thorn, and Breman
(1990) recently examined the intrarater reliability for the
Test of Orientation for Rehabilitation Patients (TORP).

Two therapists trained in the administration of the TORP
served as the raters. One therapist administered the test
while the second therapist observed and scored the test.
The test was administered to 69 persons and intrarater
reliability estimates were computed with intraclass corre-
lation coefficients.

Intrarater reliability referred to the consistency of
judgments made by the same rater over a period of time.
In most cases in which a specific analysis was labeled as
intrarater reliability, the interval between ratings by the
same examiner was brief and the focus was on outcome
measures that required some observation or judgment on
the part of a rater. Gauvin, Riddle, and Rothstein (1990)
examined both the intrarater and intrarater reliability of a
modified version of the fingertip-to-floor method of as-
sessing forward bending (lumbar flexibility). Pairs of ran-
domly determined therapists took four measurements on
73 patients with low back pain. The intrarater and intrar-
er reliability of the measurements were determined
with the intraclass correlation approach.

In contrast to intrarater reliability, the term test-retest
reliability, defined below, was usually used to describe
studies that used a longer interval between examinations
and in which the outcome was assessed by a measuring
device or instrument that did not require a rater observa-
tion or judgment.

Test-retest Reliability. Studies that were concerned
with the consistency of an evaluation or test score over
time were placed in this category. In the typical test-retest
investigation, subjects were assessed at one point in time
and reassessed after a predetermined length of time. The
two scores for each administration were then statistically
compared. For example, Burnett, Betts, and King (1990)
recently examined the test-retest reliability of measure-
ments of hip muscle torque obtained from the Cybex II
isokinetic dynamometer1. Twenty-nine male subjects
were tested two times with the Cybex II at 1- to 2-week
intervals. Intraclass correlation coefficients were used to
compute the test-retest reliability for four different mus-
cle groups.

Internal Consistency (Reliability). Studies in this
category were concerned with the consistency of items or
subparts of a particular test or assessment device. For
instance, Katz, Itzkovich, Averbuch, and Elazar (1989) re-
ported the results of a study examining the reliability of
the Loewenstein Occupational Therapy Cognitive Assess-
ment (LOTCA) Battery for patients with brain injury. The
LOTCA contains 20 subtests and is divided into four areas.
The authors reported internal consistency (reliability) co-
efficients for each of the four areas. These internal consis-
tency coefficients (coefficient alphas) were obtained by
correlating individual subtest scores within particular
areas of the assessment battery.

Procedures

The database included 20 studies examining at least one
type of reliability referred to above. The 20 studies were
obtained from a review of AJOT and Physical Therapy.
Beginning with the last issue in the 1991 volume and
working backward, we examined each issue until we
found 10 reliability studies from each journal. The range
of years searched for AJOT was 1987 through 1991; for
Physical Therapy, 1989 through 1991. All studies includ-
ed at least one assessment of reliability. In addition, the
studies had to provide the statistical value associated with
the reliability assessment and relevant descriptive infor-

1Manufactured by Cybex, division of Lumex, Inc., 2100 Seiptitown Ave-

ue, Ronkonkoma, New York 11779.
mation including sample size, number of raters, number of test items or subtests, and interval between test and retest when relevant.

Many of the 20 reliability studies included more than one type of reliability assessment. It was not uncommon for a single study to report interrater, intrarater, or internal consistency values. The Katz et al. (1989) study, for instance, included reliability values for interrater and internal consistency. In addition, studies often included more than one reliability value for a particular category. For example, in the Deitz et al. (1990) article, interrater reliability values were provided for subjects with and subjects without brain injuries. Interrater reliability values were also provided for the total score as well as for the five domains included in the TORP. Thus, the unit of analysis for examining reliability was not the study itself, but the individual reliability comparison. The 20 reliability studies included a total of 215 individual reliability coefficients. A listing of the 20 studies included in the analysis is contained in the Appendix.

Information from each of the 20 studies was recorded with a specially developed coding form. The information recorded included the year of publication, author names, title, source of publication, sample size, classification of reliability, nature of the study (i.e., outcome investigation or instrument development), the type of reliability coefficient computed, the exact reliability value, probability values, number of raters or number of trials or both, and length between evaluations (when appropriate).

Reliability of Coding

All studies were coded by the second author, a research assistant with rehabilitation experience and graduate level course work in statistics and research methods. Ten of the 20 studies were randomly selected and coded by a second rater with research experience and doctoral level research training. Reliabilities for all coded items were computed with either Kappa or intraclass correlations depending on whether the coded information was categorical or continuous. Reliability values ranged from .86 to 1.00.

Results

The 20 reliability studies contained 215 reliability coefficients. The 10 reliability studies published in AJOT contained 69 reliability coefficients; the 10 studies published in Physical Therapy included 146 reliability coefficients. Additional descriptive information concerning the reliability coefficients reported in the two journals is presented in Table 1.

Inspection of Table 1 reveals that reliability articles in Physical Therapy contained substantially more reliability coefficients ($M = 7.50$) than articles appearing in AJOT ($M = 3.45$). The sample sizes for reliability studies published in the two journals were similar, with AJOT including slightly more subjects per study. This difference may have been related to the type of reliability studies published, that is, outcome and efficacy studies versus investigations that focused on instrument development. Outcome or efficacy studies conducted in clinical environments generally have small sample sizes (Ottenbacher & Barrett, 1990). The reliability articles from AJOT included no outcome and efficacy studies, whereas 11% of the reliability coefficients from Physical Therapy were from outcome or efficacy studies.

There were also differences in the categories of reliability found across the two journals. Sixty-eight percent of the reliability coefficients included in articles retrieved from AJOT examined interrater reliability, but only 15% of the reliability values reported in AJOT addressed intrarater or test-retest reliability. In contrast, 71% of the reliability values reported in Physical Therapy examined intrarater or test-retest reliability. This difference may be related to the type of dependent variables and assessment instruments included in the studies. Interrater reliability values are generally computed when examiners are required to make a judgment or observation concerning subject performance. Many of the assessments used in occupational therapy are based on observation and require a clinical judgment, such as activities of daily living (ADL) scales. In contrast, test-retest reliability is more frequently computed when the outcome involves a response that is recorded with a mechanical device or instrument, such as recordings from a Cybex or other electrical or mechanical device.

<table>
<thead>
<tr>
<th>Variable</th>
<th>AJOT</th>
<th>Physical Therapy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>69</td>
<td>146</td>
<td>215</td>
</tr>
<tr>
<td>%</td>
<td>32</td>
<td>68</td>
<td>100</td>
</tr>
<tr>
<td>Reliability coefficients</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interrater</td>
<td>47</td>
<td>42</td>
<td>89</td>
</tr>
<tr>
<td>Intrarater</td>
<td>0</td>
<td>33</td>
<td>48</td>
</tr>
<tr>
<td>Test-retest</td>
<td>10</td>
<td>38</td>
<td>48</td>
</tr>
<tr>
<td>Internal</td>
<td>12</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>%</td>
<td>17</td>
<td>17</td>
<td>34</td>
</tr>
<tr>
<td>Nature of study</td>
<td>0</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Outcome/efficacy</td>
<td>0</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Instrument development</td>
<td>69</td>
<td>100</td>
<td>169</td>
</tr>
<tr>
<td>Type of reliability</td>
<td></td>
<td></td>
<td>93</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>37</td>
<td>54</td>
<td>91</td>
</tr>
<tr>
<td>Pearson product-moment</td>
<td>12</td>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>Correlation</td>
<td>17</td>
<td>14</td>
<td>31</td>
</tr>
<tr>
<td>Kappa</td>
<td>2</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Percent agreement</td>
<td>3</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>13</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>%</td>
<td>4</td>
<td>10</td>
<td>14</td>
</tr>
</tbody>
</table>

$^a M = 47.87, SD = 26.91. ^b M = 37.23, SD = 29.24. ^c M = 42.56, SD = 27.89.$

Table 1
Reliability Coefficients Reported in AJOT and Physical Therapy
The intraclass correlation coefficient (ICC) was the most commonly reported reliability index in both journals, accounting for 57% of all reliability coefficients. The Pearson product-moment correlation (r) was the second most frequently reported measure of reliability, accounting for 15% of all values reported across the two journals. Percentage agreement and Kappa values were infrequently reported in the articles reviewed from AJOT. In Physical Therapy the coefficient alpha was not reported in any study. This finding is not surprising, given that none of the studies reviewed in Physical Therapy were categorized as concerned with internal consistency (see Table 1).

The average values for the five identified reliability coefficients are presented in Figure 1. Examination of the figure reveals that the average value for Kappa is considerably below the average for the other four reliability procedures. All four reliability procedures have a ceiling value of 1.00 or 100%.

Discussion

Kerlinger (1986) argued that to be interpretable an assessment device or outcome measure must produce information that is reliable. The importance of developing reliable instruments and assessment procedures in applied rehabilitation fields such as occupational therapy is widely recognized. For example, Johnston, Findley, De-Luca, and Katz (1991) recently observed that “reviews of measurement in rehabilitation have noted that the field greatly needs more formal study of the reliability and validity of its measurement and assessment procedures” (p. 114). Unfortunately, numerous investigators have noted considerable confusion in the applied research literature concerning the design, analysis, and interpretation of reliability studies. Along this line, Knapp (1985) described several areas of conceptual confusion regarding reliability, agreement, and validity and provided examples from the nursing literature.

Another area of concern is the method of data analysis. Numerous authorities have commented on the inappropriate statistical analysis found in many reliability studies (Dunn, 1989; Knapp, 1985). Of specific concern is the frequent use of Pearson product-moment correlations (r) to compute reliability values. Suen and Ary (1989) suggested that the relatively convoluted concept and the corresponding convoluted theory behind the use of the Pearson product-moment correlation as an estimate of reliability has led to considerable confusion in the applied literature. The confusion is most evident when product-moment correlation values are used as a measure of interrater, intrarater, or test-retest reliability (agreement). The data from two hypothetical raters illustrate the problem of using correlation values to estimate interrater agreement (see Table 2 and Figure 2). The product-moment correlation values for both sets of data are 1.00, indicating perfect reliability; however, the agreement between the two raters is not consistent across the two data sets. The agreement between the two raters for the second data set (B) is very poor. The two raters do not agree on any score for a single subject. This inconsistency occurs because the Pearson product-moment correlation measures linear association or covariation between values, not agreement. As long as the raters co-vary consistently, a large actual disagreement in rater scores does not lower the reliability (see Figure 2).

To avoid the covariation problem associated with the Pearson product-moment correlation, and also to assess

![Figure 1. Mean reliability values from 20 therapeutic research articles.](http://ajot.aota.org/)

Table 2

<table>
<thead>
<tr>
<th>Subject</th>
<th>Data Set A Scores</th>
<th>Data Set B Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rater 1</td>
<td>Rater 2</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: r = 1.00 for both data sets.
reliability of categorical measures, researchers frequently use a simple method of computing percentage agreement among raters (Kelly, 1977). Percentage agreement methods, however, have been criticized for not taking chance agreement into account. Their use as an index of agreement among raters has been discouraged (Suen & Ary, 1989). Bartko and Carpenter (1976) provided an excellent discussion of the quantitative inadequacies of several methods commonly used to compute reliability, including Pearson product-moment correlation, percentage agreement, and chi-square. Bartko and Carpenter (1976) argued that the intraclass correlation approach (ICC) and Kappa are the preferred methods of computing reliability in applied environments. Numerous authorities have advocated the use of these two methods, particularly the intraclass correlation approach based on generalizability theory (Dunn, 1989; Kerlinger, 1986; Suen & Ary, 1989).

The intraclass correlation approach uses analysis of variance (ANOVA) to estimate true variance and error variance associated with measurement. The advantages of the intraclass approach are that it can provide estimates of both association and agreement and that it can be used with more than two sets of data (raters, trials, etc.). Shrout and Fleiss (1979) identified six different versions of ANOVA that may be used to compute intraclass correlation coefficients. The researcher using the ICC must decide whether a one-way or two-way ANOVA will be used, whether a random or fixed effects design will be used, and which variance components will be included in the computation of the ICC. When these decisions have been correctly made, the ICC provides a powerful and flexible method of examining both association (reliability) and agreement. For example, the data presented in Table 2 were analyzed with the ICC approach and the results appear in Table 3. A random effects two-way ANOVA was used to compute the variance estimates. Inspection of the results for Data Set A reveals that all of the variance is associated with the subjects. The variance associated with the raters and subject by rater interaction is zero. In this case the results of the Pearson product-moment correlation and the ICC analysis agree. The results for Data Set B, however, reveal substantial variance associated with the factor of raters. This variance reflects the poor agreement between the two raters (see Table 2 and Figure 2). When the variance components for raters are included in computing the ICC values for Data Set A and Data Set B, the ICC values are 1.00 and 0.49 respectively. If the variance components for raters are not included in computing the ICC values, then the ICC produces values that are essentially the same as those of the Pearson product-moment correlation. For detailed information concerning the appropriate ANOVA model to use in computing the ICC, the reader should consult Bartko and Carpenter (1976), Berk (1979), Dunn (1989), or Suen and Ary (1989).

The intraclass correlation approach is most frequently used with continuous data. Both Kappa and the percentage agreement method, in contrast, are used to compute reliability for categorical data or data involving dichotomous decisions. In Figure 1, the Kappa values are considerably lower than the values reported for the percentage agreement index. The percentage agreement index is

Table 3
| Variance Components for ICC Analysis for Hypothetical Data From Two Raters |
|--------------------------|------------------|
| Source                   | MS               | ICC Value |
| Data Set A               |                  |           |
| Subjects                 | 1,833            | 1.00      |
| Raters                   | 0                |           |
| Subjects by Raters       | 0                |           |
| Data Set B               |                  |           |
| Subjects                 | 4,125            | 0.49      |
| Raters                   | 15,125           |           |
| Subjects by Raters       | 458              |           |

Note. Original data presented in Table 2.
dex does not account for chance agreements (Rae, 1988). The discrepancy between the average Kappa value and the average for the percentage agreement indexes illustrated in Figure 1 may occur, in part, because Kappa corrects for chance agreement. Indirect evidence for this interpretation was provided by Suen and Lee (1985). After reanalysis of a sample of published data with the percentage agreement method, they found that if Kappa had been used, 50% to 75% of the data would have been judged as having unacceptably low reliability. That is, in examinations of categorical or dichotomous data, the percentage agreement method produced values that were inflated due to chance agreement.

Study Limitations and Conclusions
It is encouraging that most of the reliability coefficients reported in both AJOT and Physical Therapy were derived with the intraclass correlation approach (i.e., 57%). The next most frequently computed statistical procedure was the Pearson product-moment correlation. The use of this product-moment correlation and percentage agreement indexes accounted for 22% of the total reliability indexes computed across the two journals (see Table 1).

The sample of studies included in this investigation was relatively small (N = 20), although the number of reliability coefficients analyzed was adequate (N = 215). The sample of studies selected included only those in which reliability or validity and validity were specifically identified as the primary focus of the study. This practice restricted the sample to those studies concerned with instrument development and assessment. Only 7% of the total reliability coefficients reported were derived from efficacy or outcome investigations. Many studies investigating therapeutic outcomes include information on the reliability of the data collection procedures and dependent variables. The primary focus of these investigations, however, is on reliability. The reliability analysis is generally a secondary concern. These studies were not well represented in the sample selected and future research should examine the nature of reliability assessment in efficacy and outcome studies to determine similarities and differences with the results reported in this investigation. For example, many studies using single-subject designs contain individualized outcome measures. Kelly (1977) and Mitchell (1979) reported that most studies using single-subject designs in psychology or the behavioral sciences used percentage of agreement indexes as the preferred measure of interrater reliability.

Another limitation of the present study was its exclusive focus on the methodology and analysis of reliability assessment. Considerable confusion continues to exist regarding the concepts of reliability, validity, agreement, and sensitivity in applied measurement. In some cases, investigators examining the reliability of a measure would be better served by reconceptualizing the question and examining differences between means or standard errors of measurement. Altman and colleagues (Altman, 1992; Altman & Bland, 1983) have proposed a method they call the limits of agreement. This approach focuses on graphing mean differences between raters (or tests) along with standard errors and may provide data that are more useful in clinical decision making than standard reliability coefficients. The use of new approaches and the correct application of existing methods requires a clear understanding of both the methodological and conceptual issues associated with reliability assessment. An understanding of reliability and agreement in relation to measurement in occupational therapy will assist clinicians in establishing the "practice claims of the profession" (Gillette, 1982, p. 499).

Acknowledgment
Kenneth J. Ostenbacher, M.D. OTR, was supported by a Field Initiated Research Grant from the U.S. Department of Education, National Institute on Disability and Rehabilitation Research (#H135600183) during the period in which this study was completed.

Appendix
Reliability Studies Used in Analysis

Katz, N., Yacovitch, M., Averbuch, S., & Elazar, B. (1989). Loewenstein Occupational Therapy Cognitive Assessment (LOTCA) battery for brain-


## References


