A Breakdown of Reliability Coefficients by Test Type and Reliability Method, and the Clinical Implications of Low Reliability

RICHARD A. CHARTER
Department of Veterans Affairs
VA Long Beach Healthcare System

ABSTRACT. The author presented descriptive statistics for 937 reliability coefficients for various reliability methods (e.g., alpha) and test types (e.g., intelligence). He compared the average reliability coefficients with the reliability standards that are suggested by experts and found that most average reliabilities were less than ideal. Correlations showed that over the past several decades there has been neither a rise nor a decline in the value of internal consistency, retest, or interjudge reliability coefficients. Of the internal consistency approaches, there has been an increase in the use of coefficient alpha, whereas use of the split-half method has decreased over time. Decision analysis and true-score confidence intervals showed how low reliability can result in clinical decision errors.

Key words: clinical decision errors, reliability, true scores, true-score confidence intervals

DO SOME TEST SCORES, such as personality or intelligence, have higher reliabilities than do other test scores? Do some reliability methods, such as coefficient alpha or retest, provide higher reliability coefficients than do other reliability methods? Do the average reliability coefficients for scores that are used in making decisions about examinees meet the reliability standards suggested by the experts? How does low reliability affect clinical decision making? In the present article, I shall answer these questions and others concerning the impact of low reliability by presenting data from reliability studies found in test manuals, journals, and books of test critiques.

There are three general reliability categories under classical true-score theory: internal consistency, retest, and interjudge. The most important relia-
bility for the interpretation of an examinee’s score is the internal consistency reliability. Coefficient alpha, split-half, alternate forms, and KR-20, which produces the same value as alpha but is used for dichotomous data, are the most common internal consistency approaches. The *internal consistency coefficient* is a measure of the “here-and-now, on-the-spot” reliability. It indicates how close the examinee’s obtained score would come to the true score if the test were a perfect measuring instrument (Charter & Feldt, 2002). Unfortunately, the coefficients ignore the errors of measurement that are associated with day-to-day changes in examinee motivation, health, and alertness, and those errors can be significant. *Retest (stability) coefficients* are a measure of how well test scores hold up over time. We expect traits such as intelligence to hold up well over time and have relatively high retest coefficients, whereas we expect states such as depression to fluctuate over time and have relatively low retest coefficients. *Interjudge (interscorer, interrater) reliability* is a measure of the closeness of judges’ scores. It is important to know the interjudge reliability when one uses qualitative scoring (e.g., answers to essay questions) to measure an examinee’s ability or achievement.

Over the years, a number of researchers have proposed minimum levels of internal consistency reliability that a test score must have to still be useful. For example, Kelly (1927) argued that a reliability coefficient of .94 was needed to evaluate levels of individual accomplishment. That standard was widely accepted by academicians but rarely met in practice until the 1950s (Guilford, 1950). Gregory (1999), Guilford and Fruchter (1978), Hopkins, Stanley, and Hopkins (1990), and Salvia and Ysseldyke (1988) suggested a reliability of .90 for accurate measurement in practical use. For situations in which tests are clinically applied, Nunnally (1978) stated, “In such instances, it is frightening to think that any measurement error is permitted” (p. 246). Nunnally and Bernstein (1994) proposed that for research purposes (i.e., exploring the difference between groups), a reliability of .80 is adequate, but they went on to state, “If important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard” (p. 265). Aiken (1991), Rosenthal and Rosnow (1991), and Weiner and Stewart (1984) suggested a reliability of .85 or higher when scores are used to make clinical decisions. Sternberg (1994) stated, “For diagnostic and screening tests one will be very uncomfortable with reliability estimates below .80 and prefer the reliability to be .90 or higher” (p. 954). Kline (2000) stated, “Reliabilities should ideally be high around .9, especially for ability tests. Certainly alpha should never drop below .7” (p. 13). Cicchetti (1994) suggested the following reliability \( r \) guidelines for clinical significance: \( r < .70 \) (unacceptable), \( .70 \leq r < .80 \) (fair), \( .80 \leq r < .90 \) (good), and \( r \geq .90 \) (excellent).

It is difficult to establish standards for retest reliability because many factors need to be considered, such as the time between pretest and posttest, learning obtained from the pretest or between testings, and the type of test (trait or state).
Some researchers believe that it is sufficient to know that the retest coefficient is statistically significant from zero, although Huck and Cormier (2001) have warned against such use.

Cicchetti (1994) presented the standard for interjudge reliability when he suggested the following guidelines for clinical significance: \( r < .40 \) (*poor*), .40 \( \leq r < .60 \) (*fair*), .60 \( \leq r < .75 \) (*good*), and \( \geq .75 \) (*excellent*). For the one- and two-judge case, Charter (2003b) investigated the ratio of the standard deviation of the examinees’ mean difference score to the standard deviation of the differences between the judges’ ratings. He concluded that Cicchetti’s standards are too lenient, and he suggested that interjudge reliability should be held to a higher standard than that for internal consistency reliability. For example, if one uses Cicchetti’s guideline of .40 for “fair” reliability, then one finds that the standard deviation of the judges’ score differences is 1.55 times larger than the standard deviation of the subjects’ score differences. Cicchetti’s lower limit to “excellent” clinical significance—a reliability of .75—corresponds in the two-judge case to having the standard deviation of the judges’ score differences being equal to the standard deviation of the subjects’ score differences, but it is difficult to call that an “excellent” situation. To complicate matters, the degree of consistency among the judges affects the reliability. The consistency factor makes it possible for two independent groups of examinees to have exactly the same judge scores but considerably different interjudge reliabilities. Charter’s study (2003b) was preliminary and offered no standard for interjudge reliability except to say that a single standard was not sufficient. There needs to be a standard for consistent judges, a standard for inconsistent judges, and standards for those judges who cannot be categorized as either consistent or inconsistent.

Hogan, Benjamin, and Brezinski (2000) studied the frequency of the use of a variety of reliability methods. They obtained their data from a different source from that used in the present study: the *Directory of Unpublished Experimental Mental Measures, Volume 7* (Goldman, Mitchel, & Egelson, 1997). That source provided information for 2,078 tests found in 37 journals that were primarily in the fields of education, psychology, and sociology for the years 1991 to 1995. They looked at every third test and found data on 696 reliability values and reliability methods. In the present study, I surveyed several test sources other than the one used by Hogan et al., and I categorized the coefficients by reliability methods and test types. I compared the coefficients of reliability methods and test types with reliability standards, and I analyzed the coefficients to determine whether their values had changed over recent decades. In the present article, I stress the importance of reliability by showing the effects of poor reliability on clinical decision making. In my analysis, I treated test scores as a measure of the examiner’s true score (an absolute usage) rather than as a score that was used to explore individual differences (a relative usage).

Reliability is an important property of any test score. However, teaching the
importance of measurement is not a priority in most educational institutions (e.g., Aiken et al., 1990; Meier & Davis, 1990). Pedhazur and Schmelkin (1991) stated, “Measurement is the Achilles’ heel of sociobehavioral research. Although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned” (p. 2). At my institution, which is a large general medical, surgical, psychiatric, nursing home, and hospice medical center, I searched the archives of clinical psychology intern records and found that from 1980 to 2002 only 19% (27 out of 140, or 1 out of 5) had a graduate class in measurement. It appears that other outlets, such as this journal, will be largely responsible for teaching the importance of reliability to the majority of those who use test scores as part of their assessment procedures.

Method

I gathered the data from 2,733 test critiques, eight journal articles (Charter, Adkins, Alekoumbides, & Seacat, 1987; Charter, Walden, & Hoffman, 2000; Charter & Webster, 1997; Folbrecht, Charter, Walden, & Dobbs, 1999; Lopez, Charter, & Newman, 2000; Lopez, Charter, & Schelling, 2000; Rapport, Charter, Dutra, Farchione, & Kingsley, 1997; Shaw, 1966), and 47 test manuals. From those sources, I found a total of 937 reliability coefficients. I found the test critiques in Andrulis (1977), Corcoran and Fisher (2000), Hersen and Bellack (1988), Johnson (Vols. 1–2, 1976), and Keyser and Sweetland (Vols. 1–9, 1984–1992). Andrulis evaluates and describes instruments that measure adult behaviors, including aptitude, achievement, personality, attitude, creativity, and vocational, managerial, and personal performance. The purpose of the Corcoran and Fisher book is “to provide practitioners and students with a number of instruments that can be used to help them monitor and evaluate their practice” (p. iii). They selected tests that measure most of the common problems seen in clinical practice. The tests are short, easy to administer and score, and can be used for couples, families, children, and adults with problems outside the family. Hersen and Bellack’s book provides a “quick reference guide for researchers and clinicians alike needed for the multitude of assessment devices used in widely diverse problems and disorders in children, adolescents, adult, and geriatric populations” (p. xix). Johnson’s two volumes were “designed to keep researchers, evaluators, and clinicians up-to-date in child behavior” (p. ix). They contain descriptions of intelligence, readiness, maturation, achievement, cognitive style, language and number skills, anxiety, perceptions of environment, self-concept, attitude, interest, and social behavior tests. Keyser and Sweetland’s nine volumes contain lengthy descriptions on a wide range of tests: child, intelligence, marriage, family, neuropsychology, personality, research, educational (achievement, aptitude, academic subjects, school readiness, development and special education), counseling (for those in counseling), teacher
evaluation, business and industry (clerical, computer, interests, interpersonal skills, management, supervision, mechanical abilities, and manual dexterity). Each test critique included information about reliability, validity, and normative development.

Approximately one third of the 2,733 test critiques contained data that was sufficient enough to be used in this study. The data gathered from all sources were (a) the type of reliability coefficient (e.g., alpha), (b) the type of test (e.g., personality), (c) the year of the reliability study, and (d) the value of the reliability coefficient. If the precise date of the reliability study was not provided, then the date of the manual was used as the study date. The dates ranged from 1927 to 2001; 92.5% of data was from the years 1960 to 1990. There were nine instances in which the date of the study could not be established. For the 84 interjudge reliability coefficients, the sources did not state the method that was used to compute the coefficients. Therefore, the interjudge coefficient could be, for example, one of the many intraclass coefficients, Pearson’s correlation, or kappa. I analyzed only interjudge coefficients and not percentage agreement data.

**Results**

Table 1 shows the descriptive statistics for several reliability methods and the total sample. The internal consistency category was for those instances in which the source did not state the exact method of internal consistency used. If one combines the first five categories in Table 1, then one finds a mean internal consistency reliability of .82 ($SD = .130$), the sample size is 368, and the values for the 25th, 50th, and 75th percentiles are .76, .85, and .92, respectively. Therefore, the reliability that is most likely to be used in practical application to an individual test score has, on average, an internal consistency reliability of .82, which is lower than that suggested by most experts.

Table 2 shows a breakdown of the internal consistency reliability coefficients for 10 test types. The percentiles provide some indication of the percentage of coefficients that met the standards suggested by the experts. For example, the vocational and attitude tests have 25% of the reliability coefficients meeting the standard of .90.

Tables 3 and 4 show the breakdown for retest and interjudge reliability coefficients by type of test. In Table 4, the total sample provides the best description for the interjudge reliability approach because the sample sizes were small.

I did not attempt to do a statistical analysis of the mean reliability values because it was not clear that it would be meaningful to state that one type of reliability method or test type would produce, on average, higher coefficients than would other reliability methods or test types.

There were statistically nonsignificant correlations between the date of the reliability study and the internal consistency measure, $r(362) = .009$, $p = .86$, retest, $r(436) = -.041$, $p = .39$, and interjudge coefficients, $r(80) = -.026$, $p = .82$. 
### TABLE 1. Descriptive Statistics for Reliability Coefficients for a Variety of Reliability Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>M</th>
<th>SD</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>min</th>
<th>max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>KR-20</td>
<td>.83</td>
<td>.10</td>
<td>.76</td>
<td>.85</td>
<td>.91</td>
<td>.61</td>
<td>.98</td>
<td>62</td>
</tr>
<tr>
<td>Alpha</td>
<td>.81</td>
<td>.16</td>
<td>.74</td>
<td>.84</td>
<td>.93</td>
<td>.23</td>
<td>.98</td>
<td>97</td>
</tr>
<tr>
<td>Alternate forms</td>
<td>.84</td>
<td>.10</td>
<td>.81</td>
<td>.85</td>
<td>.91</td>
<td>.40</td>
<td>.96</td>
<td>40</td>
</tr>
<tr>
<td>Split-half</td>
<td>.83</td>
<td>.12</td>
<td>.76</td>
<td>.87</td>
<td>.91</td>
<td>.35</td>
<td>.98</td>
<td>126</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>.80</td>
<td>.13</td>
<td>.73</td>
<td>.83</td>
<td>.91</td>
<td>.49</td>
<td>.94</td>
<td>43</td>
</tr>
<tr>
<td>Retest</td>
<td>.79</td>
<td>.13</td>
<td>.71</td>
<td>.81</td>
<td>.89</td>
<td>.17</td>
<td>.99</td>
<td>439</td>
</tr>
<tr>
<td>Interjudge</td>
<td>.86</td>
<td>.14</td>
<td>.82</td>
<td>.90</td>
<td>.95</td>
<td>.36</td>
<td>.99</td>
<td>84</td>
</tr>
<tr>
<td>Other or unknown</td>
<td>.80</td>
<td>.16</td>
<td>.74</td>
<td>.84</td>
<td>.92</td>
<td>.15</td>
<td>.96</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td>.81</td>
<td>.13</td>
<td>.74</td>
<td>.83</td>
<td>.91</td>
<td>.15</td>
<td>.99</td>
<td>937</td>
</tr>
</tbody>
</table>

*The internal consistency category was for those instances in which the source did not state the exact method of internal consistency.*

### TABLE 2. Descriptive Statistics for Internal Consistency Reliability Coefficients for Ten Test Types

<table>
<thead>
<tr>
<th>Test</th>
<th>M</th>
<th>SD</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical adult</td>
<td>.86</td>
<td>.10</td>
<td>.82</td>
<td>.87</td>
<td>.93</td>
<td>40</td>
</tr>
<tr>
<td>Clinical child</td>
<td>.82</td>
<td>.15</td>
<td>.72</td>
<td>.87</td>
<td>.94</td>
<td>20</td>
</tr>
<tr>
<td>Aptitude</td>
<td>.85</td>
<td>.06</td>
<td>.79</td>
<td>.87</td>
<td>.89</td>
<td>12</td>
</tr>
<tr>
<td>Personality</td>
<td>.77</td>
<td>.13</td>
<td>.71</td>
<td>.78</td>
<td>.86</td>
<td>47</td>
</tr>
<tr>
<td>Vocational</td>
<td>.84</td>
<td>.09</td>
<td>.81</td>
<td>.87</td>
<td>.90</td>
<td>40</td>
</tr>
<tr>
<td>Attitude</td>
<td>.81</td>
<td>.12</td>
<td>.77</td>
<td>.85</td>
<td>.90</td>
<td>22</td>
</tr>
<tr>
<td>Neuropsychological</td>
<td>.82</td>
<td>.16</td>
<td>.79</td>
<td>.86</td>
<td>.92</td>
<td>32</td>
</tr>
<tr>
<td>Intelligence</td>
<td>.85</td>
<td>.11</td>
<td>.80</td>
<td>.86</td>
<td>.94</td>
<td>22</td>
</tr>
<tr>
<td>Educational</td>
<td>.80</td>
<td>.14</td>
<td>.80</td>
<td>.84</td>
<td>.94</td>
<td>30</td>
</tr>
<tr>
<td>Achievement</td>
<td>.91</td>
<td>.05</td>
<td>.86</td>
<td>.91</td>
<td>.95</td>
<td>12</td>
</tr>
<tr>
<td>Other</td>
<td>.81</td>
<td>.13</td>
<td>.72</td>
<td>.81</td>
<td>.94</td>
<td>91</td>
</tr>
</tbody>
</table>
TABLE 3. Descriptive Statistics for Retest Reliability Coefficients for Ten Test Types

<table>
<thead>
<tr>
<th>Test</th>
<th>M</th>
<th>SD</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical adult</td>
<td>.74</td>
<td>.18</td>
<td>.64</td>
<td>.79</td>
<td>.86</td>
<td>36</td>
</tr>
<tr>
<td>Clinical child</td>
<td>.77</td>
<td>.13</td>
<td>.73</td>
<td>.80</td>
<td>.86</td>
<td>41</td>
</tr>
<tr>
<td>Aptitude</td>
<td>.81</td>
<td>.08</td>
<td>.76</td>
<td>.81</td>
<td>.89</td>
<td>12</td>
</tr>
<tr>
<td>Personality</td>
<td>.79</td>
<td>.10</td>
<td>.71</td>
<td>.81</td>
<td>.86</td>
<td>64</td>
</tr>
<tr>
<td>Vocational</td>
<td>.76</td>
<td>.15</td>
<td>.68</td>
<td>.81</td>
<td>.85</td>
<td>46</td>
</tr>
<tr>
<td>Attitude</td>
<td>.78</td>
<td>.14</td>
<td>.67</td>
<td>.78</td>
<td>.89</td>
<td>19</td>
</tr>
<tr>
<td>Neuropsychological</td>
<td>.85</td>
<td>.11</td>
<td>.76</td>
<td>.87</td>
<td>.94</td>
<td>27</td>
</tr>
<tr>
<td>Intelligence</td>
<td>.80</td>
<td>.14</td>
<td>.71</td>
<td>.86</td>
<td>.90</td>
<td>36</td>
</tr>
<tr>
<td>Educational</td>
<td>.79</td>
<td>.13</td>
<td>.72</td>
<td>.81</td>
<td>.87</td>
<td>35</td>
</tr>
<tr>
<td>Achievement</td>
<td>.78</td>
<td>.20</td>
<td>.68</td>
<td>.86</td>
<td>.95</td>
<td>11</td>
</tr>
<tr>
<td>Other</td>
<td>.80</td>
<td>.12</td>
<td>.73</td>
<td>.81</td>
<td>.91</td>
<td>112</td>
</tr>
<tr>
<td>Total</td>
<td>.79</td>
<td>.13</td>
<td>.71</td>
<td>.81</td>
<td>.86</td>
<td>439</td>
</tr>
</tbody>
</table>

TABLE 4. Descriptive Statistics for Interjudge Reliability Coefficients for Six Test Types

<table>
<thead>
<tr>
<th>Test</th>
<th>M</th>
<th>SD</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical adult</td>
<td>.79</td>
<td>.14</td>
<td>.67</td>
<td>.83</td>
<td>.91</td>
<td>18</td>
</tr>
<tr>
<td>Clinical child</td>
<td>.84</td>
<td>.07</td>
<td>.81</td>
<td>.85</td>
<td>.90</td>
<td>11</td>
</tr>
<tr>
<td>Personality</td>
<td>.86</td>
<td>.13</td>
<td>.71</td>
<td>.92</td>
<td>.96</td>
<td>3</td>
</tr>
<tr>
<td>Neuropsychological</td>
<td>.96</td>
<td>.04</td>
<td>.92</td>
<td>.97</td>
<td>.99</td>
<td>12</td>
</tr>
<tr>
<td>Intelligence</td>
<td>.91</td>
<td>.05</td>
<td>.85</td>
<td>.93</td>
<td>.95</td>
<td>3</td>
</tr>
<tr>
<td>Educational</td>
<td>.90</td>
<td>.09</td>
<td>.81</td>
<td>.94</td>
<td>.96</td>
<td>9</td>
</tr>
<tr>
<td>Other</td>
<td>.85</td>
<td>.18</td>
<td>.82</td>
<td>.92</td>
<td>.97</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>.86</td>
<td>.14</td>
<td>.82</td>
<td>.90</td>
<td>.95</td>
<td>84</td>
</tr>
</tbody>
</table>

Discussion

Hogan et al. (2000) presented the frequency of use of different reliability methods. If one compares the frequencies from the study by Hogan et al. with those of the present study, then one will find several differences. For example, in
regard to the usage percentage of alpha (combining alpha and KR-20 and not counting the internal consistency category), Hogan et al. found 72%, whereas in the present study I found 19%; Hogan et al. found 19% were retest, whereas I found 47% were retest; Hogan et al. found 1% were interjudge; whereas I found 9% were interjudge; and Hogan et al. found 0% were alternate forms, whereas I found 4% were alternate forms. The differences may be owing to the fact that Hogan and his colleagues and I used different sources; Hogan and colleagues used unpublished tests found in 37 journals for the years 1991 to 1995 for their study, whereas I used mainly published tests of which 92.5% spanned the years 1960 to 1990. The data from both studies suggest that statements concerning the most or least popular reliability methods are premature; other sources (e.g., Educational Testing Service, 1995; Maddox, 1997) need to be examined.

Table 1 shows no obvious meaningful differences between the mean or median reliability coefficients for the various reliability methods. The largest difference was between the low of .79 for retest reliability and the high of .86 for interjudge reliability. Overall, the mean values were not low enough to be cause for concern nor were they high enough to be cause for elation. The differences between the four internal consistency methods (KR-20, alpha, alternate forms, and split-half) were slight, ranging from .81 to .84.

Table 2 also shows no obvious differences between internal consistency reliability for various test types. Personality tests had the lowest mean (.77) and median (.78) values, and achievement tests had the highest mean (.91) and median (.91) values. In general, the median values for many tests appear to be in a range (between .85 and .90) some experts would like to see for reliability coefficients that are used to interpret an examinee’s test scores, but that fall below the .90 most experts suggest.

In 1951, Cronbach referred to coefficient alpha as “a tool that we expect to become increasing prominent in the literature” (p. 299). Therefore, in the present study, the large number of split-half coefficients, which represent 35% of the internal consistency coefficients, compared with coefficient alpha (including KR-20), which represents 43% of the internal consistency coefficients, needs to be explained. The good news is that the use of the split-half method has declined over time. In the time periods before 1960, during the 1960s, the 1970s, and the 1980s, and beyond 1990, one finds that 71%, 51%, 38%, 26%, and 25%, respectively, of internal consistency methods were of the split-half type, whereas for the same time periods, 29%, 27%, 51%, 66%, and 75%, respectively, were of the coefficient alpha type. Therefore, there was a decrease in the use of the split-half method and an increase in the use of the alpha coefficient. Hogan et al. (2000) found that for the years 1991 to 1995, 94% of coefficients were alpha and 6% were split-half. It appears that Cronbach was correct.

Also in regard to the split-half method, the reporting sources rarely stated if the Spearman–Brown formula or the two-part alpha coefficient had been used to compute the reliability. Therefore, I had no indication that when the split-half
method was appropriate the best formula—the two-part alpha (Charter, 2001b; Feldt & Charter, 2003)—had been used.

There was also no indication that any of the recent advances in reliability methods published in the technical literature are being applied more generally. For example, the Angoff–Feldt or Gilmer–Feldt coefficients (Angoff, 1953; Feldt, 1975; Feldt & Brennan, 1989; Gilmer & Feldt, 1983) and latent trait approaches were not mentioned in any of the test reviews. Techniques that are used for the statistical analysis of the reliability coefficients, such as tests of statistical significance among alphas (Charter & Feldt, 1996; Feldt, Woodruff, & Salih, 1987; Woodruff & Feldt, 1986), split-half (Charter, 2001c), and intraclass coefficients (Alsawalmeh & Feldt, 1992, 1994a, 1994b, 1999), were also never mentioned. Part of the problem may be that the technical literature is too technical; perhaps researchers in that field are not familiar with technical literature outside their area of expertise. In addition, popular texts such as those by Anastasi and Urbina (1997) and Nunnally and Bernstein (1994) do not provide out-of-the-ordinary reliability coefficients and methods of statistical analysis.

The mean and median retest coefficients in Table 3 are difficult to judge because there are no retest standards and there is too little information in the present study about the retest coefficients such as the time between pretest and posttest. Still, finding that all the mean and median values are between .70 and .85 cannot be deemed poor.

If one correlates the reliability coefficients with the date of the reliability study, then it appears that from about 1960 to the present, little has changed in terms of the size of the internal consistency, retest, and interjudge reliability coefficients. Therefore, although most of the experts have suggested reliability coefficients greater than or equal to .90 as the best standard, the test constructors have not heeded the call. It may be that constructing a test to have a high reliability is just too difficult, that is, it would require too many items, which would make the test less attractive to the consumer. There is a historical precedent for such a theory (Feldt, 2001, personal communication): In 1942, Lindquist believed there was a need for achievement tests that assessed the more general goals of secondary education, which resulted in the Iowa Tests of Educational Development (ITED). When Lindquist planned the battery, he decided to make each test long enough so that it would have a reliability of .91, and he worked out that each of the main tests had to be about an hour long to achieve that. That meant that it would have taken about 7 hr, spread over 2 or 3 days, to administer the entire battery. Because the tests were unique to Iowa schools and had unmatched quality, the battery became popular in that state, and about 95% of the schools used it. In 1960, some school officials began objecting to the battery, saying that it was too long and that the children would not or could not work at maximum mental effort for 60 min. They threatened that if the tests were not shortened to the length of a standard class period of 40 to 45 min at most, then they would abandon the ITED for another test. Test
constructors were thus forced to shorten the tests—making many compromises to Lindquist’s original intentions in the process—and the tests’ reliability was lowered to .87.

Perhaps the lesson here is that it does not pay to insist that tests be built to have an internal consistency reliability of .90 or .95. Reliability is related to test length, and length is related to testing time. Tests can be built to any reliability specifications one cares to set, but publishers have a higher criterion—what users will accept and buy. If the market rejects a product, for whatever reason, it does not matter that it meets the high standards of critics and psychometric authorities. Standards are not to be scorned, and if we could meet the required standards and win consumer acceptance, then it would be great. However, typically the test users are not willing to pay the price. Some experts have set the standards so high that publishers have no alternative but to ignore them.

In regard to interjudge reliability, Cicchetti’s (1994) guidelines are quite liberal and the mean values shown in Table 4 easily meet the “excellent” category. The fact that 81% of the interjudge reliability coefficients fell into the “excellent” category, and 12% fell into the “good” category suggests that the guidelines are too lenient. Also, if the results of Charter (2003b) hold true, then the median value of .90 for the 84 interjudge coefficients is not a good sign because much higher interjudge reliabilities would be preferred.

The internal consistency reliability is the most important reliability type with which to evaluate individual examinees. It is the most appropriate coefficient for constructing confidence intervals for true scores (Charter & Feldt, 2001a). How did the internal consistency coefficients in the present study measure up to the experts’ reliability standards? The median for all the internal consistency methods was .85, which indicated that about half of the coefficients were lower than the standards recommended by some of the experts who use .85 as the standard. For those who set .90 as the standard, almost three quarters of the internal consistency coefficients fell short. On the one hand, it was encouraging to find that over 25% of the reliability coefficients were above the standard of .90. On the other hand, it was discouraging to find that 25% of the coefficients were lower than .76 and that 14% fell below .70.

The internal consistency median values for the test types shown in Table 2 indicate that more than half of the reliabilities were above .85, except for the personality tests, which had a median of only .78. Also, except for the personality tests, all the tests had more than 25% of the coefficients greater than .90. All the test means and medians met the liberal standards for clinical significance presented by Cicchetti (1994).

Most coefficients do not meet any high reliability standards, and that may bother those who would prefer tests that are used to make decisions about examinees to have high reliabilities of say, greater than or equal to .90. However, Charter and Feldt (2001b) concluded that it is not theoretically defensible to set a universal standard for test-score reliability. The aforementioned experts have given
their opinions of the “acceptable” reliability. Those opinions show considerable variability and would appear to have little or no theoretical basis. Because they are little more than personal opinions, they may not help a clinician who has to assess an examinee. It might be that the attempt to set a lower limit of acceptable reliability is a problem with no rational solution. In many clinical cases, there is no alternative evidence that is more reliable than the test score. To say that the scores are not reliable enough does not help.

That is not to say that high reliability should not be valued when clinicians have to make important decisions. For example, in a study of the meaning of reliability in terms of correct and incorrect clinical decisions (Charter & Feldt, 2001b), the following example was used. Suppose clinicians believe that an examinee who is in the bottom 10% with respect to certain abilities truly needs a special treatment. Test A has an internal consistency reliability of .90, and a short form of the same test (Test B) has a reliability of .70. An examinee below the tenth percentile on either test will need the treatment. Situations such as this will produce false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP). Will the percentages of FN and FP be smaller if Test A is used rather than Test B? If the tests have perfect reliability, then the percentage of TP, TN, FP, and FN would be 10%, 90%, 0%, 0%, respectively. With a reliability of .90, one finds 7.8% TP, 87.8% TN, 2.2% FP, and 2.2% FN. With a test score reliability of .70, one finds 6.0% TP, 86.0% TN, 4.0% FP, and 4.0% FN. Going from reliabilities of 1.0 and .90 to .70, one correctly identifies 10%, 7.8%, and 6.0%, respectively, of the examinees truly in need of treatment. In terms of decision making, the lower the reliability the more likely one is to make errors. Examiners are free to use that type of analysis to determine whether the score reliability they use is good enough, in terms of errors, for their use, or to set reliability standards. For clinicians stuck with using a test of low reliability, that means that his or her clinical decision may be tainted by measurement error. For example, at a reliability of .70 almost half the true positives will be missed.

What can examiners do when they are confronted with a test score of low reliability? It should be foremost in their minds that the obtained score, regardless of the reliability value, is a fallible score, and it should not be treated as if it were the examinee’s true score. In fact, it may be wise for examiners always to consider an obtained score as wrong because one can be almost certain that the obtained score is not equal to the examinee’s true score. Fortunately, examiners can use internal consistency reliability coefficients to construct true-score confidence intervals to express their confidence that the obtained score is close to the true score. Confidence intervals should also remind examiners of the inexactness of the obtained score and, in some instances, provide an indication of the range of possible true scores. Charter and Feldt (2001a) studied several true-score confidence interval approaches and found that the following two approaches have sound theoretical backing: (a) the traditional approach, which uses the
standard error of measurement centered on the obtained score, and (b) the regression-based approach, which uses the standard error of estimate centered on the estimated true score. If examiners construct true-score confidence intervals by using either approach, then they have an objective measure of the obtained score unreliability (measurement error); high reliability produces a narrow confidence interval and more confidence that the obtained score is close to the true score, whereas low reliability produces a relatively large confidence interval and less confidence in score interpretation. The use of confidence intervals does not, however, cancel out the effects of low reliability. For example, Charter (1996) presented a figure that showed several 95% true-score confidence intervals centered on various estimated true scores for five reliabilities. The figure’s heading was Wechsler’s classifications for IQ (Wechsler, 1981). At a reliability of .75, the confidence interval could cover the following classifications: Mentally Retarded, Borderline, Low Average, and Average. At the other end of the classification scale, the confidence interval could cover true scores from the following classifications: Average, High Average, Superior, and Very Superior. It would be difficult for a school counselor to explain to a child’s parents that the child’s intellect may be classified as either mentally retarded, borderline, low average, or average. In terms of IQs, an examinee with an IQ of 125, for a test-score reliability of .75, could have a 95% confidence interval interpretation such as “For examinees with an obtained score like yours, 95% have true scores between 106 and 132.” That is a range of 26 IQ points (almost two standard deviations), which is not encouraging or helpful in many situations.

Of course, examiners who are faced with low reliability could use more than one test. Also, in some instances the use of difference scores in the form of the “reliable difference” or the “nonoverlapping intervals” approach may be helpful (Charter & Feldt, 2000; Feldt, 1967). However, one should note that difference scores frequently have low reliabilities. Researchers have shown that an imprecise estimate of the population reliability can affect clinical decisions (Charter, 1999, 2001a, 2003a). A crucial factor in determining precision is the size of the sample (N) that is used to calculate the reliability. Whereas a few experts have made suggestions about sample size, only one study has set out to determine the ideal sample size: Charter (1999) found that at a sample size of approximately 400, there is little precision gained by adding more subjects. Therefore, not only should test constructors strive for high reliability, but they should strive for precise reliability coefficients by using 400 or more subjects whenever possible. Finally, if one holds the sample size constant, one finds that low reliability coefficients are not as precise as are high reliability coefficients.

In general, one might say that most internal consistency reliability coefficients are too low for clinicians to make precise, confident interpretations of examinee test scores. Clinicians cannot expect their tests to have perfect reliability, but they can do better. Even though it may not be absolutely necessary to have high reliability in tests, high reliability is preferable for clinicians who have to
make important decisions. The ideal of an internal consistency reliability of .90 or .95 may be difficult to attain in many cases, but test constructors should strive to give clinicians good tools to work with. In some instances, raising the reliability means simply adding a few more equivalent test items.

REFERENCES


*Manuscript received October 16, 2001
Revision accepted for publication July 23, 2002*