5-1-2005

A Meta-Analysis of Test-Retest Reliability Studies

Bryan Cromwell
Western Kentucky University

Follow this and additional works at: http://digitalcommons.wku.edu/theses
Part of the Psychology Commons

Recommended Citation
http://digitalcommons.wku.edu/theses/494

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.
A META-ANALYSIS OF TEST-RETEST RELIABILITY STUDIES

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts

By
Bryan Cromwell

May 7, 2005
A Meta Analysis of Test Retest Reliability Studies

Date Recommended: April 28, 2005

Director of Thesis: Elizabeth Shoenfelt

Dean, Graduate Studies and Research: Date

E. Gray 5/12/05
Table of Contents

Abstract ........................................................................................................ iv
Introduction ................................................................................................. 1
Method .......................................................................................................... 9
Results ........................................................................................................ 10
Discussion ................................................................................................... 17
References .................................................................................................. 21
The effect of intertest interval on test-retest reliability has received little attention in the research literature. In addition, no information has been gathered regarding the intertest intervals currently being used in test-retest reliability studies. This meta-analysis consists of 172 test-retest reliability studies. We hypothesized that the length of the intertest interval would be negatively associated with the test-retest reliability coefficient. The results of the study show that a median intertest interval of 14 days is typically used in test-retest reliability studies. Data are also provided on typical intertest intervals and test-retest reliabilities categorized by the type of test and the type of sample used in the study. Of the 172 studies used to gather normative data, 69 studies used sample sizes large enough (minimum $N = 100$) to be used in the test of our hypothesis. The results show that there is a significant negative correlation, $r = -.31$, between intertest interval and test-retest reliability. This relationship is linear when used to describe intertest intervals shorter then 90 days, but shows a nonlinear relationship when studies using intertest intervals longer then 90 days are included.
Introduction

Measurement is never perfect. Three theories of measurement, Classical Test Theory (CTT; Lord & Novick, 1968), Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), and Item Response Theory (Lord, 1980), have been proposed to model the components of a test score and how well a given measurement device performs. Of these theories CTT is the oldest, most researched, and most popular (Pedhazur & Schmelkin, 1991). CTT’s primary description of the quality of measurement is called reliability. Reliability is important because it is a necessary but not sufficient condition for quality measurement (Murphy & Davidshofer, 1998); that is, inferences drawn from scores on an unreliable test cannot be valid for any purpose, but inferences drawn from scores on a reliable test may be valid for a given purpose. In short, validity is specific to a given purpose but reliability is relevant for all purposes, and as such, it is important to determine whether a test is reliable.

According to CTT, a person’s score on a test (called the observed score) is a function of two components: the test taker’s true score, which remains constant over time, and error, which changes on every administration of the test. Thus, error in the CTT model is limited to random error (e.g., guessing), and true score is defined as the part of the observed score that is not error (literally, observed score minus random error score). If the observed scores, error scores, and true scores were known for a group of subjects, the variances of these terms could be computed. Reliability is defined as the ratio of true score variance to observed score variance. A test that yields very few random errors of measurement on a given administration has a high ratio of true score variance to observed score variance, and thus, a high reliability coefficient.
Unfortunately, a test taker's error score is never known. As a result, the test taker's true score is also never known. All that is known is the observed score. A possible solution to this problem is offered by the fact that CTT limits the error term to be random errors only. A random variable (e.g., die rolls) will have a zero correlation with any other variable. As such, if one had two versions of a test that were equal in quality in every way but contained different sets of questions, then a correlation between scores on these tests would give us the reliability coefficient. These two versions of a test are called parallel tests. Thus, CTT reliability is also defined as the correlation between parallel tests. This new definition of reliability (correlation between parallel tests) is equivalent to the earlier definition of reliability (ratio of true score variance to observed score variance). To illustrate, consider the following examples. If a test measures nothing but random error, then there will be no correlation between parallel tests and reliability will be zero, which is the same value one would have obtained with the first definition (ratio of true score variance to observed score variance) had the actual error scores been known. If there is no random error, then the scores on the parallel tests will correlate perfectly because true scores remain constant and every test taker would receive the exact same score for both forms. Again, this outcome represents the same value that one would have obtained from the measure of the ratio of true score variance to the observed score variance.

One problem with CTT relates to the assumption that all error is random. Error in measurement is of two types: systematic error and random error. Because a random variable will have a zero correlation with any other variable, CTT adequately accounts for random error. However, unaccounted for by CTT is systematic error. Systematic errors are those errors in measurement that would consistently affect an individual’s score
on the test during repeated administrations. An example of systematic error is general test
taking ability (called testwiseness). The observed scores of a group of test takers would
consistently reflect both their true scores on the test as well as their test wiseness; that is,
error due to test taking ability would remain relatively constant and upon multiple
administrations, and this ability would affect the observed score the same way. Because
CTT partitions item or test score variance into only two areas (true score or random
error), and because systematic errors are not random, systematic errors are included with
the true score term of the model. As a result, any estimation of reliability using CTT
would overestimate reliability by the inclusion of systematic error in its estimation of true
score.

*Estimating Reliability*

Although computation of a reliability coefficient appears to be easy, given that it
is defined as simply a correlation between parallel tests, in practice it is impossible to do
because parallel tests are a hypothetical entity (Crocker & Algina, 1986). It is impossible
to construct two versions of a test that consist of different items but are equal in quality in
every way. Although it is possible to construct two tests with different items that are very
close in quality, the definition demands that they be equal. In short, reliability cannot be
computed. It is, however, possible to *estimate* reliability with some strategies that satisfy
part of the definition of parallel tests. Because each of these procedures does not fully
meet the definition of parallel tests, the resultant estimates of reliability will never be
perfectly accurate, and in some cases, they will be biased (Crocker & Algina, 1986). The
three most common ways to estimate CTT reliability are: the alternate forms method, the
split half (or internal consistency) method, and the test-retest method.
In the alternate forms method for estimating reliability, two complete versions of a test are developed and given to a group of test takers. Total scores are computed for each test taker for each version of the test. The correlation between the scores serves as the estimate of reliability. The clear weakness of this approach is that although both versions of the test might be very similar to each other, they will never function equally well (i.e., same mean, same standard deviation, same correlations with other variables), and, thus, they will not be parallel. This lack of perfect equality attenuates the correlation between the scores. As a result, the resultant reliability estimate will be biased in a downward direction (i.e., lower than it should be).

The split-half (or internal consistency) method of estimating reliability has some similarities to the alternate forms method. With the split half method the researcher does not develop two complete forms of the test. Rather, the researcher splits a single version of the test into two apparently equal halves. Test takers complete both halves of the test, which are administered as a single test. Total scores are computed for each half of the test, and the scores are correlated. The resultant correlation serves as the estimate of reliability. The split-half method shares the same weakness with the alternate forms method, a lack of complete equality between the two halves of the test. The greater the mismatch in quality between the two halves of the test, the lower the reliability estimate. Other problems with the split half method, such as test length and how the split is made, exist but are easily remedied with formulae such as the Spearman-Brown prophecy formula and Cronbach's coefficient alpha.

The final method for estimating reliability is the test-retest method. With this method, reliability is estimated as a function of the test's ability to produce consistent
results from one administration to another. In a test-retest reliability study, a single version of a test is given to the test takers twice and scores from the first administration are correlated with scores from the second. This approach clearly satisfies the first part of the definition of parallel tests (equal in quality) because the exact same set of questions are used. It is also clear that the test-retest method fails to satisfy the second part of the definition of parallel tests, namely that different items must be used. This method avoids the main drawbacks of the alternate forms and split half methods because the items are the same on both administrations. Thus, there is no underestimation of reliability inherent in using different items for each administration.

Unfortunately, use of the same test for both administrations likely affects individuals' scores on the second administration. If the time interval between the two administrations is too short, there is an increased likelihood that participants will remember the test items, and therefore, their answers are likely to be more consistent than they should be. This results in an overestimate of reliability. However, if the interval between administrations is too long, participants have an increased likelihood of experiencing life events that may actually alter their standing on the construct, lowering the correlation between scores from the first and second administrations. In short, the test-retest method for estimating reliability may underestimate or overestimate reliability, depending on the length of time between the two administrations of the test (called the intertest interval). For this reason, deciding the interval of time to use between administrations of the test is critical. Ideally, there exists an optimal intertest interval for which test takers fail to remember the items from the previous administration of the test.
and during which the test takers do not change their standing on the construct or
constructs in question.

Test-Retest Reliability Guidelines

There is, however, precious little research that supports the use of any particular
length of time between administrations. As a result, most researchers follow
recommendations made without the benefit of empirical support. In addition, different
constructs may change over time more than others, which means that an appropriate
interval length for one type of testing may not be appropriate for other types of testing.

As regards the topic of intertest interval length, it is easy to find a number of
recommendations. Suggestions range from two weeks (Pedhazur & Schmelkin, 1991) to
six months (Nunnally, 1967) between test administrations. It appears that the suggestions
are based on preferences to avoid one of the two problems with intertest interval length.
Some researchers believe that the drawbacks from long intervals between tests have more
impact; thus they suggest a shorter interval (Pedhazur & Schmelkin, 1991) or recommend
that the interval not exceed six months (Anastasi, 1982). Conversely, there are
researchers who believe that the impact of a short interval is greater and who thus suggest
a longer interval (Nunnally, 1967). Unfortunately, there is little evidence to suggest that
either longer or short intervals are more effective.

One problem with all the suggested standards involves the type of test. Some
researchers suggest that relatively enduring constructs are less likely to be affected by
longer intervals than less static constructs (Cohen & Swerdlik, 1999). If this position is
correct, a universal intertest interval recommendation is likely to bias some of the test-
retest reliability estimates. Not surprisingly, there has been no research to determine which types of testing are more sensitive to longer intervals.

In the only direct study on intertest intervals, Bornstein, Rossner, and Hill (1994) executed a test-retest reliability study with varying interval lengths. The researchers administered both projective and objective measures of dependency to three groups of subjects with intertest intervals of 16, 28 or 60 weeks. The results were mixed. For the projective measure of dependency, shorter intertest intervals were associated with higher test-retest reliability coefficients; whereas, for the objective measure the highest test-retest reliability coefficients were observed with the 28 week intertest interval. Two issues further complicate the results of the study. First, the largest sample size for any of the correlations was 18. Second, the differences between correlations were never tested for significance; given the small sample sizes, these differences would likely have been nonsignificant. As a result, it is difficult to draw any clear conclusions from Bornstein et al. as regards the relationship between the intertest interval and test-retest reliability.

The Present Study

As a means for drawing conclusions about optimal intertest intervals it may be helpful to examine previous research that has used the test-retest method. Meta-analysis is a statistical technique for aggregating the results of previous research to arrive at a conclusion. Meta-analysis allows for the sampling of a large number and great variety of previous studies. By examining multiple studies one can detect large scale trends across the studies.

Because of the lack of empirical research on optimal intertest intervals and the lack of meta-analytic research concerning intertest intervals, not only are researchers
given no guidance in choosing their intertest interval they also have no pertinent
information that defines standard practice. Because of this lack of even simple normative
data, I performed a meta-analysis of intertest intervals used in test-retest reliability
studies. I tracked not only the intertest interval but also the type of test, type of sample,
and resultant reliability estimate. Additionally, the relationship between intertest intervals
and the reliability estimate were examined. My hypothesis is that intertest intervals will
be negatively associated with the reliability estimate for a given test type.

For the analysis of my hypothesis, I restricted the sample to include only studies
with a sample size of at least 100 participants to minimize the effects of sampling error
on the test-retest reliability coefficients. Ideally, the association between intertest
intervals and reliability will be nonlinear such that after a certain number of days, the
relationship is asymptotical. This critical number would then serve as the minimum
recommended intertest interval for a given test type.
Method

Data Collection

Using the PsychInfo and Silverplatter electronic databases, I reviewed test-retest reliability studies of tests published over the past 30 years. All studies published in either APA journals or The Mental Measurements Yearbook (Plake, Impara & Spies, 2005) reporting an intertest interval, sample size, and test-level reliability estimates were included in the meta-analysis. In addition, only studies involving published tests were included to reduce the confounding effect of poorly designed instruments. Tests in Print (Murphy, Impara & Plake, 2000) was used to determine which studies used published tests. For each study, I coded the following information: the intertest interval (in days), the number of subjects, the sample type, the type of test, the name of the test, and the test-level test-retest reliability estimate.

Analyses

I computed descriptive statistics for intertest intervals across all studies as well as separately by test type and sample type. For the test of my hypothesis, I correlated the intertest interval with the test-retest reliability estimate for all studies with a minimum sample size of 100. This minimum sample size requirement was used for this analysis to minimize the influence of sampling error.
Results

The PsychInfo database was used to search all APA journals over the past 30 years for test-retest reliability studies of published tests. Tests in Print (Murphy, Impara & Plake, 2000) was used to determine if the test examined in each study had been published. In addition to APA journals, studies cited in The Mental Measurements Yearbook (Plake, Impara & Spies, 2005) were also included in the study. The Mental Measurements Yearbook Volumes 9 through 16 were accessed using the Silverplatter electronic database. Studies cited in volumes earlier than Volume 9 (i.e., previous to 1985) were not searched because they were not accessible using the Silverplatter electronic database.

Of the studies found, there were 361 cases with the necessary statistics (i.e., sample size, intertest interval, and test-retest reliability). Of these 361 cases, 33 were reported in APA journals and 328 were reported in The Mental Measurements Yearbook. Of the 361 cases, 117 were excluded from analysis because they reported a range of reliability coefficients rather than a single total scale reliability coefficient. An additional 40 cases were excluded because they reported individual subscale reliabilities rather than total scale reliability. These studies were excluded from the meta-analysis on the basis that any use of subscale test-retest reliability would result in the treatment of each subscale as a unique test. Moreover, reliability coefficients would no longer be from independent samples. Another 63 cases were excluded because they reported a range of intertest intervals instead of a single intertest interval. Finally, of the remaining cases, nine studies reported multiple total score test-retest reliabilities (with a number of intertest intervals) using the same sample. Again, because each reliability
coefficient should come from an independent sample, I used data from only the first readadministration of the test (i.e., the only estimate that corresponds to a true test-retest reliability study). These exclusions resulted in 172 usable studies from the original 361.

There were two goals addressed with this study. The first is to determine current practice as regards intertest intervals (i.e., normative data). The second is to test the hypothesis that longer intertest intervals are associated with lower test-retest reliability coefficients. For a more sensitive analysis of my hypothesis, a sample size requirement of 100 participants was used. I used this minimum sample size requirement to ensure that the correlations included in the analysis were not excessively influenced by sampling error. Of the 172 studies, 69 met the sample size requirement. All 172 studies reporting the necessary data were included in determining the normative data, but only the 69 studies with the appropriate sample size were included for the analysis testing my hypothesis.

*Intertest Intervals in Practice*

I computed both mean and median values to determine the typical intertest intervals used in practice. The median intertest interval used in the 172 studies examined was 14 days. The mean intertest interval used in the studies was 55 days (SD = 122.05). The mean is much higher than the median due to the presence of a few outliers. An inspection of the data shows that only 18 cases (10% of the total sample) used intertest intervals longer than 90 days. Two cases used extremely long intertest intervals of 1047 and 730 days, which obviously had a dramatic effect on the overall mean. An intertest interval of two weeks (the median in the sample) is consistent with Pedhazur and Schmelkin's (1991) suggestion of a two week intertest interval. There were, however, a
number of studies that used intertest intervals much smaller than their recommendation. Approximately 26% of the studies included in this analysis used intertest intervals shorter than 14 days. In addition, 7% of the cases exceeded the recommendation of Anastasi (1982) stating that an intertest interval should not exceed 6 months. The study also revealed that the median sample size used in the test-retest studies observed was 70 participants. A sample size of 70 is remarkably low given Nunnally's (1967) recommendation of a minimum sample size of 300 for a reliability study, “so that this source of sampling error is a minor consideration” (p. 188). Of the 172 studies used in the meta-analysis, only 22 (13% of the total sample) used a sample greater than or equal to the 300 participants suggested by Nunnally (1967).

The data were also categorized based on the type of construct the test was designed to measure. The median intertest interval, mean intertest interval, and mean test-retest reliabilities by test type are listed in Table 1. The studies were also categorized by the age of the sample. The median intertest interval, mean intertest interval, and mean test-retest reliabilities categorized by the age of the sample are shown in Table 2.
Table 1

*Intertest Interval and Test-Retest Reliability by Test Type*

<table>
<thead>
<tr>
<th>Test Type</th>
<th>k</th>
<th>Mdn ITI</th>
<th>Mean ITI</th>
<th>Mean r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention/ADHD</td>
<td>6</td>
<td>21</td>
<td>34.50</td>
<td>.83</td>
</tr>
<tr>
<td>Behavioral</td>
<td>11</td>
<td>14</td>
<td>27.09</td>
<td>.81</td>
</tr>
<tr>
<td>Child Intelligence</td>
<td>6</td>
<td>30</td>
<td>230.50</td>
<td>.90</td>
</tr>
<tr>
<td>College Placement</td>
<td>1</td>
<td>14</td>
<td>14</td>
<td>.85</td>
</tr>
<tr>
<td>Criminal</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>.71</td>
</tr>
<tr>
<td>Eating Disorder</td>
<td>3</td>
<td>10</td>
<td>13.13</td>
<td>.85</td>
</tr>
<tr>
<td>Infant Development</td>
<td>4</td>
<td>11</td>
<td>13.50</td>
<td>.79</td>
</tr>
<tr>
<td>Intelligence/ Academic Achievement</td>
<td>8</td>
<td>25.5</td>
<td>22.63</td>
<td>.87</td>
</tr>
<tr>
<td>Language</td>
<td>11</td>
<td>14</td>
<td>26.27</td>
<td>.92</td>
</tr>
<tr>
<td>Learning Disability</td>
<td>4</td>
<td>138</td>
<td>167.50</td>
<td>.88</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>3</td>
<td>14</td>
<td>23.33</td>
<td>.78</td>
</tr>
<tr>
<td>Math</td>
<td>2</td>
<td>10</td>
<td>10</td>
<td>.96</td>
</tr>
<tr>
<td>Memory</td>
<td>5</td>
<td>30</td>
<td>140.20</td>
<td>.86</td>
</tr>
<tr>
<td>Mood</td>
<td>22</td>
<td>14</td>
<td>24.96</td>
<td>.81</td>
</tr>
<tr>
<td>Neurological</td>
<td>5</td>
<td>14</td>
<td>29.60</td>
<td>.85</td>
</tr>
<tr>
<td>Occupational</td>
<td>8</td>
<td>32.5</td>
<td>40.25</td>
<td>.82</td>
</tr>
<tr>
<td>Other</td>
<td>22</td>
<td>34</td>
<td>91.89</td>
<td>.92</td>
</tr>
<tr>
<td>Pain</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>.67</td>
</tr>
<tr>
<td>Personality</td>
<td>1</td>
<td>42</td>
<td>42</td>
<td>.77</td>
</tr>
<tr>
<td>Test Type</td>
<td>$k$</td>
<td>$Mdn$ ITI</td>
<td>Mean ITI</td>
<td>Mean $r$</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----</td>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Personal Values</td>
<td>1</td>
<td>356</td>
<td>356</td>
<td>.58</td>
</tr>
<tr>
<td>Psychopathology</td>
<td>7</td>
<td>14</td>
<td>13.57</td>
<td>.77</td>
</tr>
<tr>
<td>Reading</td>
<td>7</td>
<td>14</td>
<td>114.29</td>
<td>.80</td>
</tr>
<tr>
<td>Relationships</td>
<td>2</td>
<td>35</td>
<td>35</td>
<td>.84</td>
</tr>
<tr>
<td>Self Concept/Self Esteem</td>
<td>2</td>
<td>196.5</td>
<td>196.50</td>
<td>.81</td>
</tr>
<tr>
<td>Social</td>
<td>9</td>
<td>14</td>
<td>27.89</td>
<td>.87</td>
</tr>
<tr>
<td>Stress/Anxiety</td>
<td>7</td>
<td>14</td>
<td>14</td>
<td>.87</td>
</tr>
<tr>
<td>Substance Abuse</td>
<td>3</td>
<td>30</td>
<td>22.33</td>
<td>.86</td>
</tr>
<tr>
<td>Suicide</td>
<td>4</td>
<td>8.5</td>
<td>13</td>
<td>.79</td>
</tr>
<tr>
<td>Team Development</td>
<td>1</td>
<td>27</td>
<td>27</td>
<td>.80</td>
</tr>
<tr>
<td>Temperament</td>
<td>2</td>
<td>120</td>
<td>120</td>
<td>.72</td>
</tr>
<tr>
<td>Visual/Perception</td>
<td>2</td>
<td>13.5</td>
<td>13.50</td>
<td>.90</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>14</td>
<td>52.27</td>
<td>.83</td>
</tr>
</tbody>
</table>

_Note._ $k =$ number of studies; $ITI =$ intertest interval.
Table 2

Intertest Interval and Test-Retest Reliability by Sample Type

<table>
<thead>
<tr>
<th>Sample Type</th>
<th>k</th>
<th>Mdn ITI</th>
<th>Mean ITI</th>
<th>Mean r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infants/Toddlers</td>
<td>31</td>
<td>14</td>
<td>65.26</td>
<td>.83</td>
</tr>
<tr>
<td>Grades 3-12</td>
<td>35</td>
<td>14</td>
<td>51.34</td>
<td>.82</td>
</tr>
<tr>
<td>College Students</td>
<td>15</td>
<td>14</td>
<td>62.87</td>
<td>.79</td>
</tr>
<tr>
<td>Adults</td>
<td>34</td>
<td>15.75</td>
<td>41.54</td>
<td>.83</td>
</tr>
<tr>
<td>Seniors</td>
<td>3</td>
<td>70</td>
<td>88.67</td>
<td>.77</td>
</tr>
<tr>
<td>Undisclosed</td>
<td>54</td>
<td>14</td>
<td>56.75</td>
<td>.89</td>
</tr>
</tbody>
</table>

Note. k = number of studies; ITI = intertest interval.

As hypothesized, the correlation between the intertest interval and test-retest reliability indicated a negative relationship, $r = -.31, p < .05$. In addition, a significant, nonlinear relationship was found via polynomial regression analysis between the intertest interval and test-retest reliability, $R = .58, p < .05$. The nature of the relationship indicated a negative association between test-retest reliability and intertest intervals until the interval reached 365 days, after which the intertest interval was positively related to test-retest reliability. The nonlinear relationship was mostly due to outliers in nine studies, which used extreme intertest intervals. To examine the relationship without the effects of these outlier studies, I excluded them and re-ran the analysis. The linear correlation remained much the same as before, $r = -.39, p < .05$. Not surprisingly, there
was no longer a nonlinear association between intertest intervals and test-retest reliability. Regression analyses of this new truncated dataset indicated a regression coefficient of .002, meaning that for intertest intervals between 0 and 90 days (the range in the truncated dataset), each additional day of intertest interval is associated with a .002 reduction in the obtained test-retest reliability coefficient.
Discussion

There have been numerous suggestions by researchers concerning the length of intertest interval that should be used to reduce the effects of both participant change on the construct and participant memory of test content. Previous suggestions have been anecdotal and not based on empirical evidence or even normative information as regards what is common practice in industry. By combining published studies of test-retest reliabilities, I was able to gather data from hundreds of previous test-retest reliability studies. As a result, I am now able to inform the field about common practice. The results of the meta-analysis show that the average (median) intertest interval is 14 days for test-retest reliability studies. It was also found that there is a negative linear relationship between test-retest reliability and the intertest interval for intervals of 90 days or less. However, when studies using intertest intervals longer the 90 days were included, the relationship was nonlinear. Caution is warranted in the interpretation of this nonlinear trend, given that only a few studies were driving the results.

Another important finding in this study was the surprisingly small sample sizes used in test-retest reliability studies. The median sample size of 70 participants suggests that a large number of the studies employing the test-retest reliability method are using samples so small that the results may be influenced by sampling error. Many of the studies failed to report the number of participants that were lost due to attrition. Therefore, it is unknown whether the original sample sizes were larger but subsequently reduced due to test takers failing to participate in the second test administration. Regardless of the intended sample size, the reported samples were often small enough to be suggestive of sampling error in the resulting reliabilities, and only
12% met the sample size requirement for a reliability study suggested by Nunnally (1967).

The negative correlation found between the intertest interval and test-retest reliability was consistent with theory that suggests that a balance between participant memory and participant change on the construct is an important consideration when designing any test-retest reliability study. It appears that shorter intervals are associated with higher reliabilities, which is thought to be the result of participants' memory of test items. Conversely, longer intervals are associated with smaller reliabilities, which is thought to be the result of participant changes on the construct being measured. This change on the construct is assumed to be dependent on the type of test being administered as some constructs are more easily altered over time. Unfortunately, due to the small number of studies in each of the test type categories, I was not able to examine this relationship.

Finally, the regression coefficient ($b = .002$) illustrates the magnitude of the reduction in reliability that will result from a particular intertest interval. The .002 regression coefficient implies that with every one day increase in intertest interval one can expect a .002 reduction in the resulting reliability coefficient. Although this coefficient can be a useful tool in determining an appropriate intertest interval, the results of the meta-analysis show that the most conservative approach would be to use an intertest interval of 90 days. It is likely that an intertest interval of 90 days allows for nontrivial amounts of participant change in true score standing. Additionally, the effects of participant memory seem to be optimally reduced.
Limitations

A limitation in the findings of this study was the inconsistent reporting of test-retest reliability information. There were a large number of studies that could have been included in this study if they had reported the appropriate information. Many studies failed to report the intertest interval used between test administrations, and an even larger number of studies failed to report test level (as opposed to subscale) reliability. It is unknown whether these studies would have affected these findings, but it must be noted that the results of this study were generated from only the portion of studies that reported the appropriate information.

The negative correlation between the intertest interval and the resulting test-retest reliability coefficients is consistent with the notion that participants are more likely to alter their standing on the construct on the second administration when intertest intervals are longer. However, there is a confound in this study. Specifically, it is possible that in this database poorly constructed tests utilized longer intertest intervals and studies of more reliable tests chose to use shorter intertest intervals. This pattern would result in the same negative correlation between the intertest interval and test-retest reliability. It was my attempt to reduce the effects of such confounds by using data only from published tests. However, the magnitude of this confound in this study is unknown.

Future Research

I desired to study the affect test type has on the association between the intertest interval and reliability but was unable to do so because of limited sample sizes per test type. It would be important to examine this relationship because it is possible that certain constructs would be more easily affected by change over time, potentially affecting test-
retest reliability estimates. Finally, it would be desirable to have an examination of the relationship between the intertest interval and test-retest reliability free of confounds. An experiment in which large independent samples were given multiple administrations of the same test at different intertest intervals (similar to Bornstein et al., 1994) could be conducted to address this question.

Summary

There appears to be a number of test-retest reliability studies using sample sizes and intertest intervals that may adversely affect the resulting reliability estimates. The results of this meta-analysis show that the median intertest interval used in the studies is only 14 days, suggesting that a large number test-retest reliability studies may be heavily biased by the artifacts associated with shorter intertest intervals (e.g., item memory). In addition, despite Nunnally’s (1967) recommendation of a minimum sample size of 300 for a reliability study, the median sample size used in studies included in the meta-analysis was only 70 participants. The results of this meta-analysis suggest that an intertest interval of 90 days has many desirable properties. Therefore, researchers conducting future test-retest reliability studies should carefully weigh their intertest interval decisions based on the negative impact an improper interval can have on the results of the study.
References


Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


842-845). Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


Furlong, M., Pavelski, R., & Sandoval, J. (2001). Childhood trauma questionnaire. In B. S. Plake & J. C. Impara (Eds.) The Fourteenth mental measurements yearbook,


*Haring, T. G. (1992).** Adaptive behavior street survival skills questionnaire. In J. J. Kramer & J. C. Conoley (Eds.) *The Eleventh mental measurements yearbook,* (pp. 11-14). Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


measurements yearbook, (pp. 226-232). Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


mental measurements yearbook, (pp. 592-595). Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


measurements yearbook, (pp. 491-493). Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of
Nebraska - Lincoln.

C. Impara, & R. A. Spies (Eds.) The Fifteenth mental measurements yearbook,
(pp. 405-410). Lincoln, Nebraska: Buros Institute of Mental Measurements of The
University of Nebraska - Lincoln.

C. Impara (Eds.) The Fourteenth mental measurements yearbook, (pp. 8-12).
Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of
Nebraska - Lincoln.

Mitchell (Ed.) The Ninth mental measurements yearbook, (pp. 921-927). Lincoln,
Nebraska: Buros Institute of Mental Measurements of The University of Nebraska
- Lincoln.

Mitchell (Ed.) The Ninth mental measurements yearbook, (pp. 1059-1063).
Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of
Nebraska - Lincoln.


Mitchell (Ed.) The Ninth mental measurements yearbook, (pp. 748-751). Lincoln,
Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


yearbook, (pp. 386-389). Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


240-242). Lincoln, Nebraska: Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


Nebraska: Buros Institute of Mental Measurements of The University of Nebraska
- Lincoln.


booklet-administered minnesota multiphasic personality inventories among
primarily chemically dependent patients. *Psychological Assessment*, 2, 276-280.

The Thirteenth mental measurements yearbook, (pp. 958-959). Lincoln, Nebraska:
Buros Institute of Mental Measurements of The University of Nebraska - Lincoln.


* Study included in normative data analysis

** Study included in normative and correlational data analysis