A Coefficient Alpha for Test–Retest Data

Samuel B. Green
Arizona State University

Transient errors are caused by variations in feelings, moods, and mental states over time. If these errors are present, coefficient alpha is an inflated estimate of reliability. A true-score model is presented that incorporates transient errors for test–retest data, and a reliability estimate is derived. This estimate, referred to as the test–retest alpha, is less than coefficient alpha if transient error is present and is less susceptible to effects due to item recall than a test–retest correlation. An assumption underlying the test–retest alpha is essential tau equivalency of items. A test–retest split-half coefficient is presented as an alternative to the test–retest alpha when this assumption is violated. The test–retest alpha is the mean of all possible test–retest split-half coefficients.

Coefficient alpha (Cronbach, 1951) is probably the most popular reliability coefficient in social-science research for measures with multiple components (e.g., items; Bollen, 1989; Feldt & Brennan, 1989) and is frequently presented as the only index to support a measure’s reliability (Green & Thompson, in press). Because of its prominent role in the literature, alpha is discussed in most, if not all, psychometric texts (e.g., Allen & Yen, 1979; Crocker & Algina, 1986; Lord & Novick, 1968). In addition, a number of articles have been written discussing when coefficient alpha should and should not be applied (e.g., Cortina, 1993; Green, Lissitz, & Mulaik, 1977; Miller, 1995; Osburn, 2000; Schmitt, 1996).

A major reason why alpha is a popular reliability coefficient is that it is easy to determine. It is an internal consistency estimate and requires only a single administration of a measure rather than multiple administrations, as with test–retest or equivalent-forms reliability. In addition, researchers are not required to make any decisions about how to divide a measure into equivalent parts, as with split-half reliability.

An assumption underlying coefficient alpha is that errors among items are uncorrelated. Because measures are administered on a single occasion with coefficient alpha, it is likely that this assumption is routinely violated (Becker, 2000; Green & Hershberger, 2000; Rozeboom, 1966) and that alphas are inflated as a function of this violation (Fleishman & Benson, 1987; Guttman, 1953; Komaroff, 1997; Maxwell, 1968; Miller, 1995; Raykov, 1998; Rozeboom, 1966, 1989; Zimmerman, Zumbo, & Lalonde, 1993). Becker (2000) recently argued convincingly that transient conditions can produce correlated errors and inflated alphas. Respondents have moods, feelings, and mental states that affect their scores on a measure at a particular time, and these transient influences are likely to vary from week to week or even day to day and thus produce changes in the measure’s scores when readministered (Schmidt & Hunter, 1996, 1999). These transient errors increase estimates of the true-score variance in the computation of coefficient alpha and thus can produce inflated reliability estimates.

Becker (2000) recommended the staggered, equivalent split-half procedure as a method to estimate reliability that would correctly assign transient errors to error variance. With this approach, researchers divide their measures into equivalent halves using standard methods (e.g., Becker, 2000; Becker & Cherny, 1992; Gulliksen, 1950) and administer the two halves on different occasions. The presentations of the two half tests should be counterbalanced such that half the respondents take one split first, and the other half of the respondents take the other split first. Standard formu-
las for split-half methods (e.g., Cronbach, 1951; Rulon, 1939) can then be applied to estimate the reliability of the total measure. The advantage of this method over equivalent-forms reliability is that it does not require the development of a second measure that is comparable to the original one. Because of the difficulty of creating an equivalent form, equivalent-forms reliability is infrequently used unless the form has a purpose other than the assessment of reliability, for example, to avoid carryover effects in the evaluation of change (Becker, 2000).

Alternatively, researchers may wish to collect test–retest data when concerned with transient error. With these data, an item covariance matrix can be computed that includes Time 1 and Time 2 item variances and a number of types of covariances, including same-time/different-item covariances, different-time/same-item covariances, and different-time/different-item covariances. The various kinds of variances and covariances are shown in Figure 1. In this article, a true-score model with transient errors is presented to account for these item variances and covariances. For this model, a new estimate of reliability is derived. This estimate may be conceptualized as a reformulation of coefficient alpha and therefore is referred to as test–retest alpha. The test–retest alpha estimates true-score variance based on the different-time/different-item covariances, whereas coefficient alpha estimates true-score variance based on the same-time/different-item covariances. To the extent that transient error is present, the test–retest alpha should be less than the coefficient alpha. It is also shown that the test–retest alpha differs from the test–retest correlation, the standard statistic for test–retest data. With test–retest correlations, estimates of true-score variance are affected not only by different-time/different-item covariances but also by different-time/same-item covariances. These latter covariances are likely to create an inflated estimate of reliability to the extent that respondents remember how they responded at Time 1 and respond similarly at Time 2. Finally, a test–retest split-half coefficient is defined, and the test–retest alpha is shown to be equal to the mean of all possible test–retest split-half coefficients.

The reliability statistics discussed in the article are illustrated using a data set presented in the Appendix. The data are test–retest item scores from 40 respondents on a four-item dispositional coping measure of emotional expression and are from a larger data set that was collected by Stanton, Kirk, Cameron, and Danoff-Burg (2000). The time between testings was 4 weeks. The item covariance matrix for these data is presented in Table 1.

Coefficient alpha and split-half coefficients are discussed prior to developing the test–retest alpha and its properties. For simplicity, time is not explicitly introduced into the notation in the section on coefficient alpha and split-half coefficients because these coeffi-
Covariance Matrix for the Emotional Expression Scale

<table>
<thead>
<tr>
<th></th>
<th>Test Retest</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.769</td>
<td>.524</td>
<td>.395</td>
<td>.295</td>
<td>.810</td>
<td>.395</td>
<td>.508</td>
<td>.323</td>
<td>.410</td>
</tr>
<tr>
<td>2</td>
<td>.524</td>
<td>.972</td>
<td>.662</td>
<td>.641</td>
<td>.728</td>
<td>.508</td>
<td>.922</td>
<td>.472</td>
<td>.628</td>
</tr>
<tr>
<td>3</td>
<td>.395</td>
<td>.662</td>
<td>.708</td>
<td>.462</td>
<td>.549</td>
<td>.704</td>
<td>.421</td>
<td>.526</td>
<td>.820</td>
</tr>
<tr>
<td>4</td>
<td>.295</td>
<td>.641</td>
<td>.462</td>
<td>.821</td>
<td>.595</td>
<td>.599</td>
<td>.364</td>
<td>.500</td>
<td>.763</td>
</tr>
</tbody>
</table>

Note. The Likert items for the scale are the following: 1. "I let my feelings come out freely."; 2. "I take time to express my emotions."; 3. "I allow myself to express my emotions."; 4. "I feel free to express my emotions." See Figure 1 for descriptive terms for the variances and covariances in this matrix.

Coefficient alpha’s ease in use comes at a price: the requirement that item data meet a number of restrictive assumptions (see Crocker & Algina, 1986, pp. 105–113, for a discussion of these assumptions). The assumptions are that a score $x_{ij}$ on an item $j$ for examinee $i$ is a function of a true score ($t_{ij}$) and an error score ($e_{ij}$),

$$x_{ij} = t_{ij} + e_{ij},$$

(3)

and that for a population of examinees, the mean of the error scores for an item is zero, the error scores for an item are uncorrelated with the true scores, and the error scores for any one item are uncorrelated with the error scores for any other item. Novick and Lewis (1967) used the term essential tau equivalency to indicate that observed scores for items must be a sum of true and error scores, and as indicated by the modifier essential, that the true scores may vary by an additive constant, $c_j$, from item to item around an arbitrary true score, $t_i^0$:

$$t_{ij} = t_i^0 + c_j.$$

(4)

If the assumption of essential tau equivalency among items is violated such that the true scores are weighted differently for different items or that different true scores underlie different items (i.e., a multidimensional measure), coefficient alpha is a lower-bound estimate of reliability.

A split-half coefficient may be considered as an alternative to coefficient alpha if the assumption of essential tau equivalency among items is not met. With this approach, items are divided into two half tests that are as equivalent as possible and then a
standard formula may be applied to the half-test scores to obtain the reliability for the total scale. Perhaps the most popular method involves computing a correlation between half-test scores and then applying the Spearman–Brown prophesy formula to estimate the scale’s reliability. However, this approach requires the two halves to be parallel. Rulon’s (1939) estimate requires only that the two halves be essentially tau equivalent (Lord & Novick, 1968). Rulon’s estimate of reliability yields results identical to those obtained from Equation 2, where $x_i$ and $x_j$ are not items but two components or half tests, denoted $H$ and $H'$:

$$\hat{\alpha}_{HH} = \frac{4\hat{\delta}_{HH}}{\hat{\sigma}_X^2}.$$  \hspace{1cm} (5)

If items are essentially tau equivalent, various split-half coefficients using Equation 5 should vary only as a function of sampling error (Osburn, 2000). However, if items are not essentially tau equivalent, split-half coefficients may vary across different splits of tests, and therefore researchers should be prudent in creating half tests.

An interesting property of coefficient alpha is that it is the mean of all possible split-half coefficients, where each split-half coefficient is computed using Rulon’s (1939) method, specifically,

$$\hat{\alpha} = \bar{\alpha}_{HH}.$$  \hspace{1cm} (6)

If the essential tau-equivalency assumption is met at the item level and all split-half coefficients vary only as a function of sampling error, these coefficients should be averaged; coefficient alpha is a simple method for computing this mean. In contrast, to the extent that this assumption is not met, coefficient alpha yields a lower-bound estimate of reliability and an equivalent split-half coefficient is preferable (Cronbach, 1951; Osburn, 2000).

In Table 2, the split-half coefficients are presented for the three possible splits of the four-item coping measure. They do not vary dramatically (i.e., .84 to .88). Consequently, for all practical purposes, it is reasonable to present the mean of these coefficients, .86, which is the value of coefficient alpha.

### Table 2: Reliabilities for the Emotional Expression Scale

<table>
<thead>
<tr>
<th>Items in each split</th>
<th>$\hat{\alpha}_{HH}$ at Time 1</th>
<th>Items in each split</th>
<th>$\hat{\alpha}_{HH}$ at Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split 1 Split 2 Time 1</td>
<td>Time 2</td>
<td>Split 1 Split 2 Time 1</td>
<td>Time 2</td>
</tr>
<tr>
<td>1,2 3,4</td>
<td>.864</td>
<td>1,2 3,4</td>
<td>.761</td>
</tr>
<tr>
<td>1,3 2,4</td>
<td>.842</td>
<td>1,3 2,4</td>
<td>.606</td>
</tr>
<tr>
<td>1,4 2,3</td>
<td>.877</td>
<td>1,4 2,3</td>
<td>.753</td>
</tr>
<tr>
<td>2,3 1,4</td>
<td>.712</td>
<td>2,3 1,4</td>
<td>.712</td>
</tr>
<tr>
<td>2,4 1,3</td>
<td>.853</td>
<td>2,4 1,3</td>
<td>.853</td>
</tr>
<tr>
<td>3,4 1,2</td>
<td>.728</td>
<td>3,4 1,2</td>
<td>.728</td>
</tr>
</tbody>
</table>

*Note.* As shown: The alpha value at Time 1 was .861, and the test-retest alpha was .765. Not shown: The alpha value at Time 2 was .931, and the test-retest correlation was .752.

(See Green & Hershberger, 2000, for a discussion of this point.) However, as discussed in the introduction to this article, researchers have increasingly argued that alpha may be inflated if the uncorrelated-errors assumption is violated. To understand better why alpha may be inflated with correlated errors, Equation 2,

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \sigma^2_{x_i}}{\hat{\sigma}_X^2},$$

is examined again. The numerator of alpha is solely a function of true-score variance if all assumptions are met. However, if error scores for items are positively correlated, the magnitude of the numerator would increase, and coefficient alpha would be inflated.

Errors may be correlated for a number of reasons. For example, correlated errors may occur if items are in structural groupings (Feldt & Brennan, 1989). Examples might include (a) comprehension items that assess understanding of a common paragraph, (b) vocabulary items that require respondents to match them to a single list of definitions, or (c) a series of items concerning a graph. Alternatively, errors may be correlated for particular types of attitude or personality scales. In particular, if items assess a single construct, are phrased very similarly, and are presented consecutively, respondents may try to answer items based partially on their responses to earlier items, including their error scores, in an attempt to demonstrate consistency in their responses (Green & Hershberger, 2000). As described earlier, Becker (2000) suggested that transient errors may produce correlated errors and inflated alphas. Because we restrict our focus in this
article to correlated errors due to transient errors, it is important to have a more in-depth understanding of the psychological processes underlying these errors.

According to Schmidt and Hunter (1999), transient errors are a function of mood, feelings, mental attitudes, and cognitive energies and vary randomly from occasion to occasion; however, on any one occasion, transient errors affect all items identically. Because at any point in time, transient errors have the same effect on all items but differ among respondents, they produce positively correlated errors and inflated alphas. Schmidt and Hunter (1996) discussed the measurement of organizational commitment to illustrate the effect of transient errors. In their example, they presented a hypothetical situation where individuals felt more excited and a greater commitment to their organization if they felt energetic but felt a lower commitment if they were sick or felt sluggish. On any particular day, some individuals would report their organizational commitment as lower on items if they had colds or hangovers, but others would report it as higher if they felt physically healthy. The effects of physical well being on organizational commitment would be transient and potentially change from day to day. Most organizational researchers would likely view these transient effects as measurement error and not as part of the construct of organizational commitment. Accordingly, they would not want to compute coefficient alpha in that it fails to treat these effects as error.

Feldt and Brennan (1989) stated “that alpha ignores day-to-day fluctuation in human performance, a major source of error in many instances” (p. 113). Becker (2000) presented results on measures of personality constructs (e.g., aggression, anger, and self-esteem) and concluded that psychological measures vary in their susceptibility to transient error. Schmidt and Hunter (1996, 1999) argued that transient error is less of a problem for some types of measures (e.g., aptitude tests and job performance ratings) but indicated that further research is required to determine the degree that various measures are susceptible to these errors.

**Test–Retest True-Score Model**

A test–retest true-score model that separates transient errors from item-specific measurement errors is presented next and pictorially in Figure 2. With this model, item scores $x_{ijk}$ for individual $i$ on item $j$ at

![Figure 2. Test–retest true-score model.](#)
time \( k \) are affected not only by true scores \( t_{ij} \) and item-specific errors \( e_{ijk} \) but also by transient errors \( e_{ik} \):

\[
x_{ijk} = t_{ij} + e_{ijk} + e_{ik}.
\]  

(7)

As indicated by the subscripts, it is assumed that (a) item true scores for an individual do not change across time, (b) different items at each point in time are affected by different item-specific errors for an individual, and (c) all items at a point in time are affected by a common transient error for an individual.

Transient errors on different testing occasions are assumed to be uncorrelated with each other and with true scores and item-specific error scores. Transient errors affect items for a particular occasion such that individuals respond more similarly to items within an occasion than between occasions. More specifically, transient error creates covariance between items within an occasion but not between items for different occasions.

Transient error variances as well as item-specific error variances are assumed to be equal across testings. A consequence of this assumption is that the variances of the test and the retest are the same. In addition, it is assumed that the item-specific errors are uncorrelated with the true scores, and the item-specific errors are uncorrelated with themselves, with one exception. Because respondents are likely to remember to some extent how they have answered items on the measure on the first occasion, they may respond similarly to the same items on retest. To take into account this phenomenon, item-specific measurement errors for any one item are allowed to co-vary across occasions.

A scale score can be computed for an individual at a particular point in time by summing the item scores, \( X_{ik} = \sum_j x_{ijk} \). Similarly, a scale true score can be calculated for an individual by summing item true scores, which do not vary across time, \( T_i = \sum_j t_{ij} \). Given that item true scores can vary from item to item by a constant (i.e., Equation 4), a scale true score for an individual at any point in time can be restated as \( T_i = \sum_j t_{ij} = \sum_j (t_{ij}^* + c_j) = Jt_i^* + C \), where \( C \) is the sum of the item constants.

For the remainder of this article, the focus is on the reliability of the scale scores as a function of item variances and covariances. Variances and covariances for item true scores are unaffected by constants, allowing true scores to differ across items for an individual. Consequently, without loss of generality, true scores will be treated as equivalent across items.

Test–Retest Alpha

The reliability of scale scores can be defined as the squared correlation between the true score for a scale and the scale score, which is equal to the ratio of the variance of the true scores for a scale to the variance of the scale scores:

\[
\rho_{\text{TF}} = \frac{\sigma_T^2}{\sigma_X^2}.
\]

(8)

These variances are the same for the test and the retest given the test–retest true-score model holds, and therefore no subscript for time is included on them. In this section, the numerator is reexpressed in terms of item true-score variances. Then, estimates are presented for the numerator and the denominator of Equation 8, assuming the test–retest true-score model holds. The resulting coefficient is the test–retest alpha.

The variance of the true score for a scale (for the test or the retest) is the sum of the variances of the item true scores plus the sum of the covariances between item true scores, that is,

\[
\sigma_T^2 = \sum_j \sigma_{t_j}^2 + \sum_j \sum_{j'} \sigma_{t_jt_{j'}}.
\]

(9)

Because the item true scores are equivalent for all items, the item true-score variances are the same for all items, and the covariance between true scores for different items is equal to the item true-score variance. Hence,

\[
\sigma_T^2 = J\sigma_{t_j}^2 + (J - 1)J\sigma_{t_j}^2 = J\sigma_{t_j}^2 + (J^2 - J)\sigma_{t_j}^2 = J^2\sigma_{t_j}^2.
\]

(10)

If the item true-score variance can be estimated, the true-score variance for a scale can be estimated as

\[
\hat{\sigma}_T^2 = J^2\hat{\sigma}_j^2.
\]

(11)

Next, I show that any one of the different-time/different-item covariances is equal to the item true-score variance.
score variance, and therefore, these covariances can be used to estimate the scale true-score variance. A different-time/different-item covariance is equal to the sum of covariances of the components of the item scores, where the components are defined by the test–retest model:

\[ \alpha_{xj2} = \alpha_{xj1} + \sigma_{xj1}^2 + \sigma_{xj2}^2 + \sigma_{xj2}^2 + \sigma_{xj1}^2 + \sigma_{xj2}^2 + \sigma_{xj2}^2 \]  

\[ + \sigma_{xj2}^2 + \sigma_{xj2}^2 + \sigma_{xj2}^2 + \sigma_{xj2}^2 \]  

\[ \alpha_{xj2}^2 = \sigma_{xj2}^2 \]  

(12)

All terms after the first term on the right side of the equation are equal to zero because it is assumed that true scores are uncorrelated with specific-error and transient-error scores, and transient-error scores and specific-error scores for an item at Time 1 are uncorrelated with transient-error scores and specific-error scores for other items at Time 2. Also, because the true score is the same for all items and time points, the first term is the item true-score variance

\[ \sigma_{xj2}^2 = \sigma_{xj2}^2 \]  

(13)

The various sample different-time/different-item covariances can be pooled to obtain an estimate of the item true-score variance, namely,

\[ \hat{\sigma}_{xj}^2 = \frac{\sum \sum \hat{\sigma}_{xj}^2}{J(J-1)} = \hat{\sigma}_{xj2}^2 \]  

(14)

This estimate of item true-score variance can be substituted into Equation 11 to yield an estimate of the scale true-score variance:

\[ \hat{\sigma}_T^2 = \frac{J\hat{\sigma}_{xj2}^2}{J-1} \]  

(15)

Equation 15 provides an estimate of the numerator of the reliability coefficient and next an estimate of the denominator is sought. The variances of the scale scores are equal across testings for the test–retest true-score model. The common-scale variance can then be estimated by taking the geometric mean of the sample-scale variances for the test and the retest or the product of their standard deviations:

\[ \hat{\sigma}_X^2 = \sqrt{\hat{\sigma}_{x1}^2 \hat{\sigma}_{x2}^2} = \hat{\sigma}_{x1} \hat{\sigma}_{x2} \]  

(16)

The numerator and denominator (Equations 15 and 16, respectively) of the expression for reliability (Equation 7) can now be estimated as

\[ \hat{\alpha}_{x1x2} = \frac{J\hat{\sigma}_{xj2}^2}{\hat{\sigma}_{x1} \hat{\sigma}_{x2}} \]  

(17)

and the resulting coefficient is the test–retest alpha.

The test–retest coefficient alpha is an accurate estimate of reliability only if the test–retest true-score model holds for item data. The model specifies true scores as unchanging (within a constant) from test to retest. The test–retest coefficient alpha is an underestimate of reliability to the extent that true scores for individuals are unstable over time such that some respondents show increases, whereas others show decreases in true scores. The same requirement is made of other reliability estimates that involve assessing respondents on two occasions, including test–retest and equivalent-forms reliability estimates. In addition, items for the test and the retest must be essentially tau equivalent; of major concern, the relationships between the true scores and item scores must be the same for all items at Times 1 and 2. If this assumption is violated, the test–retest alpha is a lower bound estimate of reliability. Also, the item-specific errors must be uncorrelated, except those between an item at Time 1 and the same item at Time 2. Potentially respondents’ answers to any one item on the first administration might have an effect not only on that same item but also on other items on the second administration. The test–retest alpha would be inflated if the correlation between errors was positive between different items administered at different times.

2 The arithmetic mean of test variance and retest variance could be used rather than the geometric mean as an estimator of the observed-score variance. A small simulation was conducted to evaluate these two estimators, assuming the test–retest true-score model holds. The population reliability of a measure was set at .50, .70, or .90, and the sample size was specified as 50, 70, or 90. On the basis of 20,000 replicates for each of the nine combinations of reliability and sample size, the arithmetic mean appeared relatively unbiased, and the geometric mean was negatively biased for the smaller sample sizes and lower reliabilities. The geometric mean showed the greatest negative bias (0.87%) when reliability was .50 and sample size was 50 (e.g., expected value of the geometric mean with a population reliability of .50, .70, or .90). The geometric mean, however, showed less variability than the arithmetic mean across replicates for all nine conditions. The mean-squared error (deviation of sample estimates from the population variance) was computed to assess overall precision (i.e., both bias and relative efficiency). For all nine conditions, the geometric mean had a smaller mean-squared error. Consequently, the bias of the geometric mean was offset by its smaller variance. On the basis of these results, the geometric mean of test and retest variances is the preferred estimator.
Test–Retest Alpha Versus Coefficient Alpha and the Test–Retest Correlation

It is important to note that two standard methods to estimate reliability—coefficient alpha and the test–retest coefficient—yield poor estimates of reliability assuming the test–retest true-score model is correct. Both methods can yield inflated estimates, as discussed in the next two sections.

Coefficient Alpha

Coefficient alpha can be defined as

\[ \hat{\alpha}_k = \frac{\hat{\sigma}^2_{x_k}}{\hat{\sigma}^2_{x_k}} \]  

(18)

to allow for different values across time, \( k \). Under the assumptions invoked to derive coefficient alpha, the same-time/different-item covariance, \( \hat{\sigma}^2_{x_k} \), is a sample estimate of the item true-score variance. However, under the test–retest true-score model, any one covariance between different items within an occasion is equal to the item true-score variance plus the transient-error variance for that occasion:

\[ \hat{\sigma}^2_{x_k} = \hat{\sigma}^2_{x_k} + \hat{\sigma}^2_{x_k} + \hat{\sigma}^2_{x_k} + \hat{\sigma}^2_{x_k} + \hat{\sigma}^2_{x_k} + \hat{\sigma}^2_{x_k} \]

(19)

Accordingly, \( \hat{\sigma}^2_{x_k} \) would on the average overestimate the item true-score variance. To the extent that transient error is present, coefficient alpha is an inflated estimate of the reliability.

As shown in Table 2, coefficient alpha for the Emotional Expression scale was .86 for the test and .93 for the retest, whereas the test–retest alpha was .74. Assuming the test–retest true-score model holds for the example, the results suggest that the scale is susceptible to transient error as reflected by the lower value of the test–retest alpha in comparison with the alphas. On the other hand, if the test–retest true-score model does not hold because respondents’ true scores changed between scale administrations, the test–retest alpha could be an underestimate of the reliability. Because emotional expression is a trait measure that should be relatively invariant between administrations (Stanton et al., 2000), the lower test–retest alpha may be interpreted as due to transient error rather than true-score change.

Alternatively, the choice between interpretations could be made on the basis of how the test is used to make decisions (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). If changes in scores across time lead to inaccurate decisions about respondents, then these changes should be interpreted as measurement errors; otherwise, they should not. For example, a psychologist Sarah may want to determine which individuals may attend a workshop based on their scores on the Emotional Expression scale. Ideally, she would administer the scale the day before the workshop to individuals and select those with the lowest scores. However, practically she is forced to administer the scale to different individuals on different days in the month prior to the workshop. Although some individuals are assessed 4 weeks prior to the workshop, and others are assessed the day before the workshop, she is comfortable comparing their scale scores because the Emotional Expression scale is supposed to be a trait measure. For this scenario, changes in the scale scores due to time should be considered measurement error (i.e., transient error) in that these changes lead to invalid decisions. Sarah would be better informed by assessing the scale’s reliability using test–retest alpha rather than coefficient alpha.

Test–Reest Correlation

The traditional reliability estimate for test–retest data is the correlation between the test and the retest, namely,

\[ \hat{\rho}_{x_1x_2} = \frac{\hat{\sigma}_{x_1x_2}}{\hat{\sigma}_{x_1}\hat{\sigma}_{x_2}} \]

(20)

The covariance between \( x_1 \) and \( x_2 \) is equal to the covariance between the component parts for \( x_1 \) and \( x_2 \). Consequently, Equation 20 can be reexpressed as

\[ \hat{\rho}_{x_1x_2} = \frac{\sum f_{j\neq x_j}\hat{\sigma}_{x_1x_2} + \sum f_{j\neq x_j}\hat{\sigma}_{x_1x_2}}{\hat{\sigma}_{x_1}\hat{\sigma}_{x_2}} \]

(21)

Because each of the sums of covariances in Equation 21 is equal to the number of covariances times their mean,

\[ \hat{\rho}_{x_1x_2} = \frac{J(J-1)\hat{\sigma}_{x_1x_2} + \sum f_{j\neq x_j}\hat{\sigma}_{x_1x_2}}{\hat{\sigma}_{x_1}\hat{\sigma}_{x_2}} \]

\[ \hat{\rho}_{x_1x_2} = \frac{J\hat{\sigma}_{x_1x_2} + \sum f_{j\neq x_j}\hat{\sigma}_{x_1x_2}}{\hat{\sigma}_{x_1}\hat{\sigma}_{x_2}} \]

(22)

The first term on the right side of Equation 22 is the test–retest alpha, and therefore

\[ \hat{\rho}_{x_1x_2} = \hat{\alpha}_{x_1x_2} + \frac{J(\hat{\sigma}_{x_1x_2} - \hat{\sigma}_{x_1x_2})}{\hat{\sigma}_{x_1}\hat{\sigma}_{x_2}} \]

(23)
The test–retest correlation is equal to the test–retest alpha and a function of the difference between the mean of different-time/same-item covariances and the mean of different-time/different-item covariances. Assuming the test–retest true-score model is appropriate, the test–retest correlation is a biased estimate of reliability to the extent that these two mean covariances are on the average not equal to each other. As previously shown, a different-time/different-item covariance is equal to the item true-score variance for the test–retest true-score model. In comparison, the different-time/same-item covariance is equal to the item true-score variance plus the covariance between the item specific-error scores for the test and the retest:

\[
\sigma_{tj\mid t_2} = \sigma_{rj\mid t_1} + \sigma_{rj\mid t_2} + \sigma_{ej\mid t_1} + \sigma_{ej\mid t_2} + \sigma_{ej\mid r_1}
\]

From a slightly different perspective, a commonly stated problem associated with the test–retest correlation is that it is inflated to the extent that responses to items on the first taking of a test affect responses to the same items upon retesting (e.g., Allen & Yen, 1979; Crocker & Algina, 1986; Lord & Novick, 1968). The different-time/same-item covariances are sensitive to these inflationary effects and intuitively might be eliminated from the test–retest correlation to correct for this bias. Thus, the test–retest alpha, which substitutes different-time/different-item covariances for different-time/same-item covariances, could be considered a test–retest correlation corrected for item memory effects.

The difference between the test–retest correlation and the test–retest alpha should be less to the extent that the test–retest correlation is not strongly influenced by the different-time/same-item covariances. Because the proportion of different-time covariances that are different-time/same-item covariances (\(J/\sum \) or \(1/J\); see Figure 1) becomes smaller as the number of items increases, the difference between the two reliability estimates should generally become smaller for longer measures.

As shown in Table 2, the test–retest correlation was .75, and the test–retest alpha was .74. The difference between the coefficients is small, particularly given the Emotional Expression scale consists of only four items. It appears as if the respondents did not respond to any marked degree to items on the second occasion on the basis of their responses on the first occasion. The two coefficients might have differed more dramatically if the time between testings had been less than 4 weeks.

**Test–Re...**
components (two halves) squared (i.e., $2^2 = 4$). The denominator, the variance of the total score, is computed by summing the half tests administered at Times 1 and 2 and calculating a sample variance on these summed scores ($\hat{\sigma}_T^2$). This coefficient is the same as the one suggested by Becker (2000) for the staggered equivalent split-half procedure.

For test–retest data, the estimate of the total score variance $\hat{\sigma}_T^2$ ignores half of the data collected at each time. An alternative estimate that takes into account all of the data is the geometric mean of the total test variance at Time 1 and the total test variance at Time 2,

$$\hat{\sigma}_X = \sqrt{\hat{\sigma}_{X1}^2 \hat{\sigma}_{X2}^2} = \hat{\sigma}_{X1} \hat{\sigma}_{X2}. \quad (26)$$

Using this estimate for the variance of the scale scores, the test–retest split-half coefficient is estimated using the statistic

$$\hat{\alpha}_{H1H2} = \frac{4\hat{\sigma}_{H1H2}}{\hat{\sigma}_X \hat{\sigma}_{X2}}. \quad (27)$$

The number of possible ways to split a test with $J$ items between a test and a retest ($N_{s\text{plit}}$) is $J$ items taken $J/2$ at a time or $J!(J/2)!^2$. For example, with a four-item measure, the number of possible splits is six; the splits are shown in Table 2. Potentially, for any measure, researchers could compute as many as $N_{s\text{plit}}$ test–retest split-half coefficients. They could then present one or more of these coefficients to portray as accurately as possible the reliability of their measure. Given no sampling error and that all assumptions are met except equivalence of halves, all test–retest split-half coefficients would be lower-bound estimates of reliability. However, the halves that are most equivalent would produce the greatest lower-bound estimate, and in this respect, the best estimate of reliability. (See Callender & Osburn, 1977, and Osburn, 2000, for a comparable argument concerning all possible split-half coefficients.) However, in the presence of sampling error, it is unclear which of the $N_{s\text{plit}}$ coefficients is best in that larger coefficients may be due to capitalization on chance.

Instead of selecting the highest possible test–retest split-half coefficient, researchers should conduct a study prior to the collection of the test–retest data to assess how a measure should be split into equivalent halves. Becker (2000) recommended a method that (a) pairs items on the basis of their loadings on the general factor; (b) allocates items within pairs to one or the other half, taking into account item means and standard deviations as well as item content; and (c) allows for reshuffling of items to produce more similar statistics between halves. Alternatively, on the basis of the work of Callender and Osburn (1977) and Osburn (2000), the initial study might involve choosing the split that maximizes the test–retest split-half coefficient. Next, a test–retest reliability study would be conducted, and a test–retest split-half coefficient would be computed for the halves selected on the basis of the initial study. This methodology would also allow researchers to reevaluate the equivalence of the selected halves and to offer, with appropriate cautions, alternative test–retest split-half coefficients if the selected halves were not found to be equivalent in the test–retest reliability study.

The test–retest split-half methodology does not allow for an examination of the effects of item ordering on test scores. Potentially the ordering of items can have an effect on scores of noncognitive measures and the reliability of these measures (Green & Hershberger, 2000; Schurr & Henriksen, 1983). In comparison, an advantage of the staggered-equivalent split-half procedure is that it permits an evaluation of item orderings on item and scale scores. This advantage is particularly important if a measure is administered in different orders to respondents.

**Test–Re-test Alpha as the Mean of Test–Re-test Split-half Coefficients**

As discussed earlier, coefficient alpha is the mean of all possible split-half coefficients. Similarly, the test–retest alpha is the mean of all possible test–retest split-half coefficients. Intuitively, it may not be surprising that the test–retest alpha and the test–retest split-half coefficient are related in that they are both a function of the different-time/different-item covariances. In this section, the relationship between these reliability estimates is demonstrated explicitly.

The mean of the test–retest split-half coefficients across $N_{s\text{plit}}$ splits can be expressed as

$$\overline{\alpha}_{H1H2} = \frac{4\overline{\sigma}_{H1H2}}{\hat{\sigma}_X \hat{\sigma}_{X2}}. \quad (28)$$

The mean covariance between halves (i.e., $\overline{\sigma}_{H1H2}$) can be reexpressed as a function of item covariances. The covariance between summed items for two half tests is equal to the sum of all covariances between pairs of items from the two different halves. For example, for the four-item coping measure, the covariance between half tests containing Items 1 and 2 at Time 1 and
Items 3 and 4 at Time 2 is $\hat{\sigma}_{x_j x_j} = \hat{\sigma}_{x_j x_j}^2 + \hat{\sigma}_{x_j x_{12}} + \hat{\sigma}_{x_{12} x_{12}}$. To compute the mean of the covariances between all possible split halves, (a) each different-time/different-item covariance is multiplied by the probability that this covariance between items is part of the covariances between half tests, among all possible ways of splitting a measure between a test and a retest and (b) the resulting set of crossproducts is then summed. The probability of the occurrence of any particular different-time/different-item covariance is $J/[4(J - 1)]$. Accordingly,

$$\overline{\sigma}_{H_1 H_2} = \sum_{j, j' \neq j}^J \frac{\hat{\sigma}_{x_j x_j}}{4(J - 1)} = \frac{J}{4(J - 1)} \sum_{j, j' \neq j}^J \hat{\sigma}_{x_j x_j}.$$  

(29)

However, the sum of the different-time/different-item covariances in Equation 29 is equal to the number of these covariances times their mean, namely,

$$\sum_{j, j' \neq j}^J \hat{\sigma}_{x_j x_j} = J(J - 1)\overline{\sigma}_{x_j x_j}.$$  

(30)

The quantity on the right of Equation 30 can be substituted for the sum of the item covariances in Equation 29 to yield

$$\overline{\sigma}_{H_1 H_2} = \frac{J}{4(J - 1)} J(J - 1)\overline{\sigma}_{x_j x_j} = \frac{J^2 \overline{\sigma}_{x_j x_j}}{4}.$$  

(31)

Substituting the right-hand expression of Equation 31 for mean of all split-half coefficients in Equation 28, the mean test–retest split-half coefficient is

$$\overline{\sigma}_{H_1 H_2} = \frac{J^2 \overline{\sigma}_{x_j x_j}}{4} \frac{\hat{\sigma}_{x_j x_{12}}}{\hat{\sigma}_{x_j x_j}}.$$  

(32)

The right-hand expression in Equation 32 is the test–retest alpha as shown in Equation 17, and consequently, the test–retest alpha is equal to the mean of all possible test–retest split-half coefficients:

$$\hat{\sigma}_{x_j x_{12}} = \overline{\sigma}_{H_1 H_2}.$$  

(33)

**Choice Between a Test–Reæst Alpha and Test–Reæst Split-Half Coefficient**

The test–retest alpha does not require a researcher to split a measure into halves and is an estimate of reliability if the test–retest true-score model is met. Presumably test developers most frequently attempt to develop scales that are unidimensional and choose items that are maximally saturated by the underlying factor. To the extent that test developers are successful, items on measures should be more or less essentially tau equivalent, and test–retest alphas should be reasonable estimates of reliability. However, if items fail to meet this requirement, a test–retest split-half coefficient is preferable.

As shown in Table 2, the test–retest split-half coefficients for the Emotional Expressiveness scale vary from .61 to .85. The mean of these coefficients is .74, which is the value for the test–retest alpha, as it must be. Given the variability in these coefficients, one might initially conjecture that the items are not essentially tau equivalent. Accordingly, the test–retest alpha of .74 would be an underestimate of the reliability. However, the lowest and highest coefficients are associated with the same split: .61 (lowest) when the halves were based on Items 1 and 3 at Time 1 and Items 2 and 4 at Time 2 and .85 (highest) when the halves were based on Items 2 and 4 at Time 1 and Items 1 and 3 at Time 2. To feel emotionally secure, one might wish to conclude that the variability in coefficients is due to sampling error and to estimate the reliability based on the test–retest alpha rather than consider some psychometrically unsavory model that would explain these perplexing results. For complex results like the ones in the example, researchers could investigate alternative models of item responding using structural equation modeling; however, they may encounter statistical problems, including nonnormality of data, insufficient sample size, and mathematical and empirical underidentification (Green & Hershberger, 2000). Given results like the ones seen here, it would be prudent to collect a large second sample to assess replicability of results and to explore the model underlying the test–retest data.

**Summary**

Social science researchers have available a large number of methods for computing reliability (see Feldt & Brennan, 1989, for a summary of many of these methods) but generally use a small subset of these methods to assess the consistency of their measures. Coefficient alpha is used most often and probably used too frequently given the assumptions underlying it (e.g., Becker, 2000; Cortina, 1993; Feldt & Brennan, 1989; Green & Hershberger, 2000; Schmitt, 1996). Becker (2000) argued that the presence of transient error violates one of these assumptions, independence of errors. He concluded on the basis of empiri-
cational analyses that transient error in item data “can range from nonexistent to very large” (p. 370). Because it is unclear what type of measures are susceptible to transient error, it can be argued that coefficient alpha, as well as other internal consistency estimates, should not be used routinely when first evaluating the psychometric quality of a measure unless they are part of a larger assessment of a measure’s reliability.

A number of reliability coefficients are available that reflect appropriately the effects of transient error, including the equivalent-forms correlation, the staggered-equivalent split-half coefficient, the test–retest split-half coefficient, the test–retest alpha, and the test–retest correlation. With these coefficients, researchers should carefully choose the time interval between administrations of measures to minimize changes in true score but allow for transient effects. These decisions can best be made if researchers understand the psychological processes underlying the measurement errors for an instrument and its application as well as the construct of interest (Schmidt & Hunter, 1999).

The equivalent-forms estimate is not likely to be applied in most cases because of the amount of work required in creating a form equivalent to the original measure. The staggered-equivalent split-half method obviates this problem. A slight revision to this method was suggested here that requires researchers to collect test–retest data and to compute a test–retest split-half coefficient. This revision allows the measure to be administered intact. Ideally researchers who wish to compute a test–retest split-half coefficient should decide on the basis of previous research how a measure can be split into equivalent halves, as is required by the staggered-equivalent split-half method. In the reliability study, researchers can examine the distribution of the test–retest split-half coefficients to ensure that the results are consistent with those from the equivalency study.

The test–retest alpha does not require a researcher to split a measure into halves and is the average of all possible test–retest split-half coefficients. However, it does require items to be essentially tau equivalent. To the extent that this requirement is not met, a test–retest split-half coefficient is preferable. Alternatively, researchers might choose to compute a test–retest correlation in that it does not require essential tau equivalency among items. However, they might also present a test–retest alpha in addition to the test–retest correlation to offer a coefficient that excludes inflationary, item-memory effects.

References


## Appendix

Sample Data for the Emotional Expression Scale

<table>
<thead>
<tr>
<th>Person</th>
<th>Test</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Re-test</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>03</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>05</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>06</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>08</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>09</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>