VALIDITY, INVALIDITY, AND RELIABILITY

David R. Heise
QUEENS COLLEGE

George W. Bohrnstedt
UNIVERSITY OF MINNESOTA

Principal support for the work in this paper came from the Office of Education (OER 5-10-822), for which the authors are grateful. They also would like to thank Arthur S. Goldberger for valuable discussions on an earlier version of this paper.

Sociologists almost always rely on fallible measuring instruments in attempting to estimate parameters. Recent work by Heise (1960) has shown how parametric estimates can be made with fallible data when one has three waves of panel data. However, when one has only cross-sectional data, he must obtain multiple measurements on the same underlying "true" variable and use the correlations among the fallible measurements to make estimates of the parameters. Two different approaches...
with multiple indicators have been taken. In the first, one usually begins with a set of items or indicators assumed to have one or more underlying causes and factor analyzes the set of items to determine the relationships of the items to the underlying factors or causes. Then the items which are the “purest” indicators (based on the factor loadings) are built into a
some kind of linear composite (for example, an index or summed rating scale). The reliability of the composite can be estimated using one of the internal consistency methods (Lord and Novick, 1968, pp. 134–137; Holzinger, 1920), and given a reliability coefficient, one can correct the obtained correlations or regression coefficients for attenuation
due to unreliability. This approach to parametric estimation with
fulfillable data has its historical roots in educational psychology and has been widely adopted by sociologists and social psychologists in recent years.

More recently, Coenen (1990) and Blalock (1990) have presented
a second approach to the problem—an approach which grows largely
out of the “errors in variables” problem in econometrics (Johnston, 1963, 
Chapter Six) and path analysis (Duncan, 1966). In this approach one
makes assumptions about the causal structure relating the items to the
underlying variables, and the underlying variables to each other, and
then, using the correlation among the indicators, obtains estimates of
the various paths.

At this time it would be premature to argue which approach is
likely to be the more fruitful. However, since the former is in wide use,
it would seem appropriate to explore it more carefully than it has been.
In this paper three issues are considered. First, what is the validity of
the composite one has created? Stated differently, what is the correlation
between the composite one has constructed and the underlying variable
which is assumed to have generated the correlations among the indicators
in the composite? The validity formula that is derived in this paper has
been obtained previously by Cattell and Radecki (1950) using a different
approach. However, their work has remained obscure. Second, the
concept of invalidity is introduced. It is possible (and often the case, one
can assume) that the indicators built into a composite show variation
due to causes (factor) other than the one they are intended to measure.
The greater the amount of this unwanted variation, the greater the
invalidity of one’s composite. Finally, the relationship between factor
analysis and the reliability of a composite score is shown, and a formula
is introduced for obtaining a reliability coefficient from factor analysis
results.

In the derivations it is assumed that the reader is familiar with
the basic procedures of path analysis (see Wright, 1934; Duncan, 1966;
Land, 1969), factor analysis (see Harman, 1967), and matrix algebra. Readers not interested in the derivations will find key formulas in the section titled, "Rescaling the Formulas," and examples of their use in the section, "Some Examples."

**DERIVATIONS**

Underlying the building of composites from factor analysis and the derivations below are a number of assumptions about how n items or indicator variables are related to one another: (1) relationships among the items are assumed to be linear; (2) the total variance in items is a function of m orthogonal latent variables, of the n disturbances that uniquely affect the item, and of m measurement errors; (3) the disturbances of the items are assumed uncorrelated with each other, with the latent variables, and with the measurement errors; similarly, measurement errors are assumed uncorrelated with each other or with latent variables; (4) the items themselves are assumed to be locally independent (Lord and Novick, 1968, pp. 360-302); that is, items are assumed to be mutually uncorrelated to each other except by their mutual dependence on latent traits; (5) the observed factor structure accurately represents the structure of the latent traits.

![Figure 1](image-url)
In the path diagram in Figure 1, T1 and T2 represent two latent variables or traits and \( r_{xy} \) represents the correlation between the traits; however, in derivations below, we will be assuming \( r_{xy} = 0 \). The \( ps \) indicate the influence of the traits on four items measured without error, labeled X1. The \( ps \) can be thought of as validity coefficients relating the items measured without error to the latent traits. The \( u \) represent unique sources of variance for the items measured without error; and the \( e \) indicate the degree of influence from unique sources. The degree of relationship between the items measured without error and the empirical items with measurement error, \( z \), is \( \sqrt{\rho_e} \) (Lord and Novick, 1968, p. 61) where \( \rho_e \) represents the reliability of the item. The extent to which observed responses are a function of errors is defined as a residual, \( \sqrt{1 - \rho_e} \). The lack of additional paths reflects various previously stated assumptions. We have shown only two latent traits and four items so as to simplify the presentation.

Following the rules of path analysis (Duncan, 1966), the diagram can be simplified somewhat as shown in Figure 2. In turn, this diagram can be translated into one employing customary factor analysis quantities as shown in Figure 2. Here, \( f_i \) is the factor loading for the ith item on the ith factor and also the validity coefficient for that empirical item, \( h_i^2 \) is the communality for that item, and the \( v \) represent the com-
biased effects of \( u \) and \( e \). The translation between Figures 2 and 3 involves the following identity in the model:

\[
J_{ik} = p_{ik} \sqrt{\rho_{ii}}
\]  

(1)

That is, the factor loading of item \( i \) on factor \( k \) equals the validity of the item measured without error times the square root of the item’s reliability.

A path diagram also can be drawn to indicate how a composite score to measure \( T \) is related to its component items (Figure 4). Here
the $\alpha_i$ represent inter-item correlations, and the $\beta_{jk}$ are path coefficients indicating how much each item affects the total score, $X_i$. It is important to notice that no disturbance term is drawn to $S_i$ since the variance of the composite score obviously is completely determined by the items.

Finally, since the latent-trait model assumes that the inter-item correlations are entirely a function of the latent traits, Figures 3 and 4 can be combined, as in Figure 5, to indicate how a composite score is linked to latent traits. Figure 5 is the key diagram for analysis in this paper. In the next section, it will be shown that the $\beta_{jk}$ are a function of the $f_j$ and of the observed inter-item correlations so all of the parameters of the model are defined in terms of statistics obtained in the course of a factor analysis. And, using the path diagram, it is possible to define the correlation between a latent trait and the composite score. For example, assuming the underlying factors are orthogonal, by the rules of path analysis, the correlation between latent trait $T_1$ and the score $S_i$ is:

$$r_{1S_i} = (f_1\alpha_{i1}) + (f_2\alpha_{i2}) + (f_3\alpha_{i3}) + (f_4\alpha_{i4})$$

(2)

The correlation between a composite $X_i$ and the variable it is meant to measure, $T_1$, is the validity of the composite. So, the path model leads to the definition of a validity index for the composite score. Note that in Figure 5 a second unwanted factor, $T_6$, also is responsible for some of the covariation among the four indicators. The procedure outlined above will later be extended to obtain indices of invalidity for such a situation. Then, reliability coefficients will also be developed from a similar procedure.
For convenience in the derivations, it is assumed that item scores are standardized, that is, have means equal to zero and standard deviations equal to unity, although formulas are eventually rescaled into the original metric of the items.

We begin by defining the matrix of inter-item correlations among items \( x_i \), \( i = 1, 2, \ldots, n \):

\[
R = \begin{pmatrix}
1 & \rho_{12} & \cdots & \rho_{1n} \\
\rho_{21} & 1 & \cdots & \rho_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{n1} & \rho_{n2} & \cdots & 1
\end{pmatrix}
\]  

(3)

Correlations among the items are assumed to be generated by mutual dependencies on factors \( T_k \), \( k = 1, 2, \ldots, m \), where \( m < n \). Throughout the derivations, the factors are assumed to be orthogonal, so both the factor pattern and the factor structure are defined in terms of a single matrix in which an element \( f_{ik} \) is the factor loading of item \( i \) on factor \( k \). The matrix of factor loadings is:

\[
F = \begin{pmatrix}
f_{11} & f_{12} & \cdots & f_{1m} \\
f_{21} & f_{22} & \cdots & f_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
f_{n1} & f_{n2} & \cdots & f_{nm}
\end{pmatrix}
\]

(4)

The loadings of items on factor \( k \) will be represented as a vector:

\[
f_k = \begin{pmatrix}
f_{1k} \\
f_{2k} \\
\vdots \\
f_{nk}
\end{pmatrix}
\]

(5)

A composite score to measure factor \( T_k \) is generated by summing items in accordance with an appropriate weighting scheme. That is:

\[
S_{ik} = \sum_{j=1}^{n} w_{ij} x_{ij}
\]

(6)

where \( S_{ik} \) is the score to measure factor \( T_k \) for observation \( g \), \( x_{ij} \) is observation of score on item \( i \), and \( w_{ij} \) is the weight assigned to item \( i \) in order to measure factor \( T_k \). The set of weights which measure one factor will be represented as a vector:
\[
W = \begin{pmatrix}
W_1 \\
W_2 \\
\vdots \\
W_n
\end{pmatrix}
\]  
(7)

Validity

Our first objective is to define the correlation between a given factor, \( T_a \), and its corresponding score, \( S_a \). Referring to Figure 6 and applying the rules of path analysis, it is seen that the correlation between factor \( T_a \) and score \( S_a \) is:

\[
\rho_{T_a S_a} = f_1 \beta_{A1} + f_2 \beta_{A2} + \cdots + f_k \beta_{Ak}
\]

(8)

Let the path coefficients linking items to a given score, \( S_a \), be represented as a vector.

\[
\beta_a = \begin{pmatrix}
\beta_{A1} \\
\beta_{A2} \\
\vdots \\
\beta_{Ak}
\end{pmatrix}
\]

(9)

Now, equation (8) can be rewritten:

\[
\rho_{T_a S_a} = \beta_a^T b_1
\]

(10)

Or, in general, the correlation between trait \( T_a \) and its corresponding score \( S_a \) is:

\[
\rho_{T_a S_a} = \beta_a^T b_a
\]

(11)

where \( b_a \) is the set of path coefficients or standardized regression weights for predicting \( S_a \) from the items, and these are defined as follows:

\[
b_a = R_{\text{item}}^{-1} r_a
\]

(12)

where \( r_a \) is the vector of item-to-total score correlations (Walker and Lev, 1953, p. 332). That is,

\[
r_a = \begin{pmatrix}
r_{1a} \\
r_{2a} \\
\vdots \\
r_{na}
\end{pmatrix}
\]  
(15)
Now consider an element $a_{ik}$ from $r_{ik}$, where $a$ refers to the $i$th item. Since the items are in standardized form, we can write from equation (6)

$$
\rho_{ik} = \frac{E(\sum_{i=1}^{n} w_{i}a_{ik}a_{ij})}{\tau_{ik}}
$$

The variance of $S_{i}$ is

$$
\sigma_{i}^{2} = E\left[\left(\sum_{i=1}^{n} w_{i}a_{ik}\right)^{2}\right] = \frac{E\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i}a_{ik}a_{ij}\right)}{	au_{ik}}
$$

Substituting the square root of equation (15) in equation (14) gives:

$$
\rho_{ik} = \frac{E\left(\sum_{i=1}^{n} w_{i}a_{ik}a_{ij}\right)}{\sqrt{E\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i}a_{ik}a_{ij}\right)}}
$$

or

$$
\rho_{ik} = \frac{\sum_{i=1}^{n} w_{i}a_{ik}a_{ij}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i}a_{ik}a_{ij}}}
$$

since $E(\sum_{i=1}^{n} w_{i}a_{ik}) = \rho_{ik}$ for standardized variables. Again, equation (17) shows the solution for one element in equation (13). The general solution for all $i$ items can be expressed in matrix form as:

$$
\tau_{n} = \frac{Rw_{i}}{\sqrt{w_{i}Rw_{i}}}
$$

The denominator in equation (18) is a scalar quantity equal simply to the square root of the sum of the weighted entries in the matrix of inter-class correlations.

Now, substituting equation (15) in equation (12), we obtain

$$
\delta_{i} = \frac{E^{-1}(Rw_{i})}{\sqrt{w_{i}Rw_{i}}} = \frac{w_{i}}{\sqrt{w_{i}Rw_{i}}}
$$

and substituting equation (19) in equation (11), we find:

$$
\rho_{i} = \frac{w_{i}}{\sqrt{w_{i}Rw_{i}}}
$$
or, in summation terms:

\[ \psi_{a,b} = \frac{\sum_{i=1}^{n} \psi_{a,b} \psi_{i}}{\sqrt{\sum_{i=1}^{n} \psi_{a,b}^2 \psi_{i}^2}} \]  

(21)

The term \( \psi_{a,b} \), as defined by the computing formulas (21), is a validity index specifying the correlation between the trait \( T_a \) and its corresponding score, \( S_a \). and \( \psi_{a,b} \) indicates the proportion of variance in \( S_a \) that is associated with \( T_a \). The use of this formula is shown in the section, "Some Examples."

**Invalidity**

The correlation between a score, \( S_a \), and some trait, \( p_a \), that \( \psi_a \) is not supposed to measure is given by a variation of equation (20):

\[ \psi_{a,b} = \frac{\sum_{i=1}^{n} \psi_{a,b} \psi_{i}}{\sum_{i=1}^{n} \psi_{a,b}^2 \psi_{i}^2} \]  

(22)

and this quantity squared indicates the proportion of variance in \( S_a \) that is due to the unwanted trait, \( T_a \). Hence,

\[ \psi_{a,b}^2 = \frac{[\sum_{i=1}^{n} \psi_{a,b} \psi_{i}]^2}{\sum_{i=1}^{n} \psi_{a,b}^2 \psi_{i}^2} = \frac{\psi_{a,b} \psi_{i}^2}{\psi_{a,b} \psi_{i}^2} \]  

(23)

When there are several unwanted traits being measured by \( S_a \), the total proportion of variance in \( S_a \) that is due to all the unwanted traits, represented below by \( \psi_{a,b} \), is simply the sum of these coefficients squared since the traits are assumed to be orthogonal. That is:

\[ \psi_{a} = \sum_{i=1}^{n} \frac{\psi_{a,b} \psi_{i}^2}{\psi_{a,b} \psi_{i}^2} \]  

(24)

Equation (24) can be rewritten as follows:

\[ \psi_{a} = \frac{w_i^T \psi \psi^T \psi \psi \psi}{\psi \psi^T \psi \psi^T \psi} \]  

(25)

The second term in equation (25) is simply the validity squared. Now, assuming that \( FF \) = \( R - I + H \), where \( I \) is the identity matrix and \( H \) is the diagonal matrix of communalities, equation (25) can be rewritten:
\[ \psi_{i}^{u} = 1 - \rho_{au} + \frac{\sum_{j=1}^{n} w_{ij}^2}{\sum_{i=1}^{n} w_{ij}^2} \]  

(20)

In summation terms, equation (20) becomes:

\[ \psi_{a} = 1 - \rho_{au} + \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} w_{ij}^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^2 \rho_{ij}} \]  

(27)

where \( \psi_{a} \) is a measure of invalidity. It indicates the proportion of variance in \( \Delta_i \) that is associated with latent variables other than the one of interest. Its actual use shall also be shown in the section, "Some Examples."

**RELIABILITY**

The reliability of a measure is defined as the correlation between two equivalent forms of a test (Lord and Novick, 1968, p. 58). Figure 6 presents a path diagram for a simplified case to indicate what this means in terms of the latent trait model. Figure 6 is like Figure 5 except, for simplicity, it deals with a single latent trait, and it shows the construction of two equivalent scores, \( S \) and \( S' \), based on different but equivalent items, \( z_i \) and \( z'_i \), \( i = 1, 2, \ldots, n \). The key point to notice in Figure 6 is that the correlation between the two equivalent scores depends only on their mutual dependence on \( T \).

Actually, it is not necessary to work out all the separate paths linking \( S \) and \( S' \) since we already have an expression for the validity.
coefficient for each test, and this coefficient is equivalent to a standardized regression coefficient or path coefficient indicating the degree of influence of \( T \) on \( S \) or \( S' \). Therefore,

\[
\rho_{ST} = \rho_{ST}^{w}\rho_{w}
\]  

(28)

However, since items in one test are equivalent to those in the other, \( \rho_{ST} = \rho_{w} \) and:

\[
\rho_{ST}^{w} = \rho_{ST}^{w} \rho_{w} \]  

(29)

However, formula (29) is not a general formula since it presumes a single factor. In the multi-factor case, \( \rho_{ST}^{w} \) would be obtained by summing the correlations between \( S \) and \( S' \) due to the different latent variables. That is (changing notations to correspond to the multiple factor situation):

\[
\rho_{ST}^{w} = \rho_{ST}^{w} + \rho_{ST}^{w} + \cdots + \rho_{ST}^{w} \]  

(30)

From equation (23) it follows that:

\[
\rho_{ST}^{w} = \sum_{i=1}^{n} \left( \frac{w_i}{w} \right) \]  

(31)

which is simply:

\[
\rho_{ST}^{w} = \frac{w_i R_i + w_i H_i - w_i}{w_i R_i} \]  

(32)

assuming \( FF' = R - I + H \). We have chosen to call formula (32) \( \Omega \). Rewritten in summation notation and rearranging terms, the reliability of composite \( k \) is:

\[
\Omega = \rho_{ST}^{w} = 1 - \frac{\sum_{i=1}^{n} w_i A_i}{\sum_{i=1}^{n} w_i A_i + \sum_{i=1}^{n} w_i A_i} \]  

(33)

Rescaling the Formulas

Traditionally, reliability formulas have been expressed in the original metric of the items. To accomplish this in our formulas, we merely substitute \( \sigma \) for \( w \), where \( \sigma \) is a vector of the \( n \) item standard deviations. Doing this, equation (31), the measure of validity, becomes
$$\varphi_{ab}^2 = \frac{\sum_{i=1}^{n} \epsilon_i \epsilon_i'}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(x_i, x_j)}}$$  \hspace{1cm} (34)$$

where it is understood that $\text{Cov}(x_i, x_j) = \epsilon_i$. Similarly, equation (27), the invalidity measure, becomes:

$$\Phi_{ab} = 1 - \varphi_{ab}^2 + \sum_{i=1}^{n} \epsilon_i \epsilon_i' - \sum_{i=1}^{n} \epsilon_i^2 \hspace{1cm} (35)$$

And, equation (33), the reliability measure, is changed to:

$$\Omega = 1 - \sum_{i=1}^{n} \epsilon_i^2 - \sum_{i=1}^{n} \epsilon_i \epsilon_i' \hspace{1cm} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(x_i, x_j)$$  \hspace{1cm} (36)$$

Note from equation (35) and (36) that $\Phi_{ab}$ can be simplified to:

$$\Phi_{ab} = \Omega - \varphi_{ab}^2 \hspace{1cm} (37)$$

**Comparisons of $\alpha$ and $\Omega$**

Novick and Lewis (1967) have demonstrated that in general $\alpha$, the most popular interval consistency measure (Cronbach, 1951), is not equal to the reliability of a composite score, but instead is a lower bound to it. It is not the only lower bound reliability estimate, but since it is the most popular, it will be compared to $\Omega$ in this section.

Cronbach defined $\alpha$ as:

$$\alpha = \frac{n}{n-1} \left[ 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(x_i, x_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(x_i, x_j)} \right] \hspace{1cm} (38)$$

Novick and Lewis (1967) prove that if the items are $r$-equivalent, that is, true scores on one item differ from true scores on another item by no more than a constant, then $\alpha$ is an exact estimate of the reliability. When one's items are essentially $r$-equivalent, $\alpha$ and $\Omega$ will equal each
other. Under this condition all the inter-item correlations are equal, and the item variances are equal, as shown in Lord and Novick (1988, p. 90). Hence,
\[ \alpha = \rho_{pp} = \frac{np}{1 - \rho - (n - 1) \rho} \tag{39} \]
where \( \rho \) is an inter-item correlation. Now, under this condition,
\[ \Omega = 1 - \frac{np^2 - np}{np^2 + n(n - 1) \rho} \\
= 1 - \frac{(1 - \rho)}{1 + (n - 1) \rho} \\
= \frac{np}{1 + (n - 1) \rho} \\
= \alpha \tag{40} \]
Thus, under the condition of essential \( r \)-equivalence, \( \alpha \) and \( \Omega \) are equal, but otherwise \( \Omega \) will generally be larger than \( \alpha \) since when the \( \bar{R} \) are known, \( \Omega \) is exactly equal to the reliability of a composite.

**Weighting Selected Items**

Frequently, when building a composite score, one assigns zero weights to those items which have low loadings on a given factor. In this case one can simply collect the items with non-zero weights and apply directly formulas (34), (35), and (36) to the reduced set of items. This is likely to apply this procedure when there are multiple factors and one wishes to include each item in but a single score. Then the low loadings of each item on the other factors are simply ignored. Of course, the values of the \( fA \) and the \( \bar{R} \) used are those obtained in the factor analysis of all items. This procedure shall be followed when the use of the formulas is shown with actual data.

If one chooses to apply weights other than the standard deviations to his items, one can use the following formulas. The validity formula (34) is changed to:
\[ \rho_{wA} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{r} w_{ij} \alpha_{ij} s_{i}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{r} w_{ij} \alpha_{ij} \text{ Cov}(fA, \bar{s}_A)}} \tag{41} \]
The invalidity formula (35) is changed to
\[
\psi_k = 1 - \rho_{kk} + \frac{\sum_{i=1}^{n} \omega_i^2 - \sum_{i=1}^{n} \delta_{ii}}{\sum_{i=1}^{n} \sum_{j \neq i} \omega_i \omega_j \text{Cov}(x_i, x_j)}
\]

and the reliability formula (30) is changed to

\[
\Omega = 1 - \frac{\sum_{i=1}^{n} \omega_i^2 - \sum_{i=1}^{n} \delta_{ii}}{\sum_{i=1}^{n} \sum_{j \neq i} \omega_i \omega_j \text{Cov}(x_i, x_j)}
\]

where \( \omega_k \) is the weight applied to the \( k \)th indicator for the \( k \) factor.

Some Examples

The data for the following examples came from an unpublished study of the values of college students. Two of the values to be measured were called religiosity and fatalism. Ones were constructed to measure these two domains and several factor analyses were done on a sample of 500 males to establish the validity of the items. The six items from which the religiosity scale was to be built and the five items for the

**Table 1**

<table>
<thead>
<tr>
<th>Items Included in the Religiosity and Fatalism Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Religiosity</td>
</tr>
<tr>
<td>1. Everyone should believe in and practice some religion.</td>
</tr>
<tr>
<td>2. The best way to foster the moral development of civilization is through organized religion.</td>
</tr>
<tr>
<td>3. There should be stricter observance of the Sabbath, the religious day of rest.</td>
</tr>
<tr>
<td>4. The world moves in an evolutionary process of unfolding rather than through divine guidance.</td>
</tr>
<tr>
<td>5. There is an almighty God who watches over us.</td>
</tr>
<tr>
<td>6. There is a life after death.</td>
</tr>
<tr>
<td>Fatalism</td>
</tr>
<tr>
<td>7. The world seems to move so slowly that the destinies of men—men do not shape the world.</td>
</tr>
<tr>
<td>8. People should live for today and let tomorrow take care of itself.</td>
</tr>
<tr>
<td>9. Since most things are inevitable, people should relax and enjoy themselves.</td>
</tr>
<tr>
<td>10. People can actually do very little to change their lives.</td>
</tr>
<tr>
<td>11. A person really has very little control over his own fate.</td>
</tr>
</tbody>
</table>

*Response categories were "strongly disagree," "probably disagree," "probably agree," and "definitely agree."*

*Weighted negatively in composite to assign the item in meaning with others in the composite.*
measure of futilism are shown in Table 1. The respondents were asked to show degree of agreement with each of the items on a four-point scale ranging from "strongly disagree" to "strongly agree."

Several methods for estimating factor loadings exist, and since these methods do not all yield exactly the same factor structure, the estimates of $\Omega$, $\psi$, and $\Phi$ will differ from one method to another. Also, one might expect somewhat different estimates of the three statistics depending upon whether one factors the two sets of items together rather than separately. In order to investigate what differences might occur in the statistics using an empirical example, the following procedures were followed. The data were factored using four different methods: a principal factor solution (Harman, 1967, Chapter Eight); two maximum likelihood solutions (Harman, 1967, Chapter Ten)—one where the squares of the multiple correlations (sas) are used as communality estimates and a second where the sas are the first values used and then iteration procedures are applied until the communality estimates from one iteration do not diverge significantly from those on the previous iteration; and, finally, the alpha procedure (Kaiser and Cuffrey, 1965) which also estimated the communality by iteration after beginning with the sas. Readers interested in the technical differences between these methods of estimating the loadings are directed to the sources indicated.

Each set of items was factored separately using the above four factoring methods, and then the two sets of items were pooled and again analyzed using the four factoring methods.

Scores were created by giving the items zero or unit weights. The covariances among the eleven items composing the two scores are shown in Table 2. These will be used in later computations.

In Table 3 we show the factor analysis of just one of the factoring methods—the maximum likelihood solution without iteration. The results of this factor analysis are typical of those of the other three methods in the following ways. First, almost without exception, when the two sets of items were factored separately, two factors emerged indicating that a score constructed from that set would contain some invalid variance. The single exception was the alpha factoring of the religiosity items which forced a single factor. Second, when the two sets were factored jointly by the four methods, three or more factors emerged, not two, again indicating some invalidity. Third, when the factors were rotated using the varimax criterion (Kaiser, 1958), the solutions made for less substantivo sense than did the unrotated solutions. Hence, all the results shown here are based on unrotated factors (that is, the principal factor solution).
<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.409</td>
<td>0.282</td>
<td>0.282</td>
<td>0.229</td>
<td>0.300</td>
<td>0.300</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>2</td>
<td>0.409</td>
<td>0.409</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
</tr>
<tr>
<td>3</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
</tr>
<tr>
<td>4</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
<td>0.300</td>
</tr>
<tr>
<td>5</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>6</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>7</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>8</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>9</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>10</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>11</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
<td>0.277</td>
</tr>
</tbody>
</table>
### TABLE 3
Factors Analysis of the Reliability and Validity Using Rotated Solution Without Rotation

<table>
<thead>
<tr>
<th>Items</th>
<th>Factor I</th>
<th>Factor II</th>
<th>Factor III</th>
<th>Factor IV</th>
<th>Rotated Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.600</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.600</td>
</tr>
<tr>
<td>2</td>
<td>0.553</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.553</td>
</tr>
<tr>
<td>3</td>
<td>0.521</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.521</td>
</tr>
<tr>
<td>4</td>
<td>0.420</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.420</td>
</tr>
<tr>
<td>5</td>
<td>0.713</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.713</td>
</tr>
<tr>
<td>6</td>
<td>0.627</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.627</td>
</tr>
<tr>
<td>7</td>
<td>0.441</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.441</td>
</tr>
<tr>
<td>8</td>
<td>0.441</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.441</td>
</tr>
<tr>
<td>9</td>
<td>0.441</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.441</td>
</tr>
<tr>
<td>10</td>
<td>0.441</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.441</td>
</tr>
</tbody>
</table>

* N = 300 college freshmen males.
### Table 4

An example of how to compute estimates of $\Omega$, $\rho_{yy}$, and $\Psi$ from the factor analysis of the religiosity items shown in Table 3.

<table>
<thead>
<tr>
<th>Item</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.599</td>
<td>0.928</td>
<td>0.472</td>
<td>0.695</td>
<td>0.563</td>
<td>0.067</td>
</tr>
<tr>
<td>2</td>
<td>0.492</td>
<td>0.486</td>
<td>0.396</td>
<td>0.511</td>
<td>0.636</td>
<td>0.344</td>
</tr>
<tr>
<td>3</td>
<td>0.459</td>
<td>0.725</td>
<td>0.305</td>
<td>0.609</td>
<td>0.392</td>
<td>0.289</td>
</tr>
<tr>
<td>4</td>
<td>0.424</td>
<td>0.894</td>
<td>0.379</td>
<td>0.639</td>
<td>0.548</td>
<td>0.404</td>
</tr>
<tr>
<td>5</td>
<td>0.655</td>
<td>0.671</td>
<td>0.419</td>
<td>0.374</td>
<td>0.519</td>
<td>0.049</td>
</tr>
<tr>
<td>6</td>
<td>0.624</td>
<td>0.964</td>
<td>0.629</td>
<td>0.773</td>
<td>0.597</td>
<td>0.771</td>
</tr>
</tbody>
</table>

**Sum**

|       | 4.990 | 2.381 | 3.817 |

$$\Omega = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij}^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(x_i, x_j)}$$

$$\rho_{yy} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} x_i x_j}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(x_i, x_j)}}$$

$$\Psi = \Omega - \rho_{yy} = 4.8165 - (0.9171)^2 = 0.0394$$

Table 4 shows an example of how to compute estimates of $\Omega$, $\rho_{yy}$, and $\Psi$ using the results from Table 3 of the maximum likelihood factor analysis of the separate set of religiosity items. Note that column 3 is the product of columns 1 and 2. The sum of column 2 and column 3 is used in computing the numerator of the second term of the expression for $\Omega$—formula (36). The denominator of that term is simply equal to the sum of all the entries in the covariance matrix for the religiosity items (upper left hand quadrant of Table 2). The numerator of $\rho_{yy}$—formula (34) is the sum of the entries in column 3, which in turn arc products of column 4 and 5. The entries in column 3 are, of course, just the square roots of the terms in the covariance matrix, which was obtained in computing $\Omega$. Finally, $\Psi$ is simply the difference between $\Omega$ and $\rho_{yy}$.

Thus, using this factoring method one would estimate the reliability to be 0.8015 and the validity 0.9171. Roughly 2 per cent of the variance in the score would be invariant, owing to the second factor.
Table 5 presents the statistics obtained from all the different methods of analysis. It is clear that estimates do depend somewhat upon the method of factoring, and whether one factors the item sets separately or together. Two findings deserve comment. In every case, no matter what factoring method was used, if estimates exceeded Cronbach’s α (the values of α for the two scores are shown at the bottom of Table 5), as one would expect since α is a lower bound to the reliability. Second, when one factors all sets of items together, the reliabilities of the composites generally increase, as expected (the single exception was for the fatality score, alpha factoring). However, note that the validity decreases and the instability increases at the same time the reliability increases. This indicates that the reliability of a composite sometimes can be increased by adding items from other domains of content, but only at the expense of validity.

Since there is no one best factoring method, it cannot be argued on a mathematical basis that any one of the estimates above is better than the other. At the same time it is to be noted that all of the procedures that were tried here gave estimates that were in the same range.

**DISCUSSION**

Some verbal elaboration of the concepts as used in this paper are in order. First, validity is defined as the correlation between a measure and the true underlying variable. A high validity coefficient does not imply that one has measured that which he set out to measure. It means only that whatever the items are measuring, the composite constructed is highly correlated to it. Other types of validity assessment also are needed as pointed out in the “Technical Recommendations” of the American Psychological Association (1960). Second, one ordinarily thinks of validity as equating the square root of the reliability, but this is not necessarily the case in practice. The validity of a composite can be considerably less than the square root of the reliability when the composite’s variance is due to several underlying factors instead of to a single factor. It may seem paradoxical that one can increase reliability by adding items from a domain of content other than that being measured. However, reliability deals only with whether individuals would have the same relative standing on measurement. Thus, a measure can be highly reliable but invalid. As shown in equation (27), the measure of invalidity is simply the difference between the reliability and the validity squared. Obviously, if this difference is zero, the validity will equal the square root of the reliability. A very small amount of algebraic manipulation of formula (27) makes it clear that the reliability of a
<table>
<thead>
<tr>
<th>Factoring Method</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Separately</td>
<td>Jointly</td>
<td>Separately</td>
<td>Jointly</td>
</tr>
<tr>
<td>A. Religiosity</td>
<td>0.8615</td>
<td>0.8223</td>
<td>0.9171</td>
<td>0.9149</td>
</tr>
<tr>
<td>Principal Components</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>0.8615</td>
<td>0.8231</td>
<td>0.9171</td>
<td>0.9171</td>
</tr>
<tr>
<td>Maximum Likelihood—Iterate</td>
<td>0.8726</td>
<td>0.8726</td>
<td>0.9309</td>
<td>0.9224</td>
</tr>
<tr>
<td>Alpha—Iterate</td>
<td>0.8570</td>
<td>0.8377</td>
<td>0.9355</td>
<td>0.9010</td>
</tr>
<tr>
<td>B. Fatalism</td>
<td>0.6177</td>
<td>0.6240</td>
<td>0.7250</td>
<td>0.7360</td>
</tr>
<tr>
<td>Principal Components</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>0.6177</td>
<td>0.6347</td>
<td>0.7230</td>
<td>0.7432</td>
</tr>
<tr>
<td>Maximum Likelihood—Iterate</td>
<td>0.6250</td>
<td>0.6025</td>
<td>0.5999</td>
<td>0.7397</td>
</tr>
<tr>
<td>Alpha—Iterate</td>
<td>0.6689</td>
<td>0.6611</td>
<td>0.6136</td>
<td>0.7559</td>
</tr>
</tbody>
</table>

$\alpha = 0.8550$ for religiosity composite.

$\alpha_f = 0.6068$ for fatalism composite.
composites is the sum of both valid and invalid variance. This fact has important consequences for correcting estimates for attenuation due to errors in measurement.

Assuming one finds more than a single factor underlying a set of items, he cannot simply correct obtained coefficients for unreliability using standard attenuation formulas because the invalid variance in one's score yields too high a reliability estimate. Instead, it is advised to use a path analytic approach such as indicated for the two-variable case in Figure 7. Here it is assumed that $x$ causes $y$ and one wants to estimate the path between $x$ and $y$, that is, $p_{xy}$. One cannot measure $x$ and $y$ directly but instead builds errors, $S_x$ and $S_y$. If one factor analyzed the two sets of items and found more than a single factor underlying each set, the causal representation of the effects on $S_x$ and $S_y$ would be that shown in Figure 7. There is reliable variance in $S_x$, due not only to $x$, but also to some other common factor(s) labeled $w$. Similarly for $S_y$, some variance is due to $w$. As was seen above, $\Omega$, the reliability estimate, is due to all common factors. Thus, simply to estimate $p_{xy}$ by dividing $s_{xy}$ by the square root of the product of the two reliability estimates would give an estimate of $p_{xy}$ which is too small. The proper approach would be to note that
\[
p_{xy} = p_{xw}p_{yw}
\]
and hence $p_{xy} = p_{xy}/(p_{xw})$.

But, $p_{xy} = s_{xy}$ and $p_{xy} = s_{xy}$—these paths simply equal the validity coefficient for $S_x$ and $S_y$. Hence, one corrects for attenuation in $p_{xy}$ not by dividing by the square root of the product of the $\Omega$s, but instead by dividing by the product of the two validity coefficients.

For example, suppose one found that $p_{xy} = 0.4$, $\Omega_x = 0.8$, $s_{xy} = 0.85$, $\Omega_x = 0.9$, and $s_{xy} = 0.92$. Using the usual correction for attenuation procedures, one would estimate that $p_{xy} = 0.4/\sqrt{(0.8)(0.9)} = 0.47$. However, the correct estimate would be $p_{xy} = 0.4/(0.85)(0.92) = 0.51$. 
The extension to the multivariate case is straightforward and, hence, is not discussed here.

The equations defining validity, invalidity, and reliability involve factor analysis, so two standard problems in factor analysis must be kept in mind when applying the formulas.

*Factor rotations* is a central concern since there are an infinite number of rotations which will satisfy any set of data (Harman, 1967, p. 24). It has been assumed here that where more than a single factor exists a unique rotation exists where the factors are aligned with latent traits, and, therefore, that factor loadings indicate the degree to which latent traits influence item scores. If the factors are markedly misaligned with latent traits, the factor loadings will not be interpretable this way, and the validity and invalidity measures will be in error. Also, the equations were derived under the assumption of orthogonal factors. Actually, the validity, invalidity, and reliability formulas all can be applied in cases of oblique factor structures, providing that pattern coefficients (as opposed to structure coefficients) are used when validity is calculated and when the formula calls for $f_{00}$. Interpretation of it is not affected by nonorthogonality. However, the validity and invalidity coefficients do shift in meaning. When traits are correlated, the validity coefficients simply specify the correlation of the same with the desired trait, and one must allow that some part of this correlation may be due to the influence of other traits that are correlated with the one of interest. The invalidity coefficient only specifies the proportion of reliable variance that is not correlated with the desired trait; it does not necessarily indicate how much variance is determined by unwanted traits. Indices for dealing more suitably with oblique factor scores are discussed by Cattell and Rudolphi (1962).

The second major problem involves the communality, or the values entered into the diagonal of the correlation matrix before factoring. Communalities have two effects on results of factor analyses: (1) they determine the number of factors extracted, and (2) they affect the size of factor loadings. Although our formulas depend upon knowing the communalities, there is no way to determine them exactly—they must be estimated. Obviously, then, one's choice of communality estimates affects the size of the validity, invalidity, and reliability coefficients presented here. There are several procedures for estimating communalities including the highest inter-item correlation a given item has with the other items, the squared multiple correlation of an item with the remaining n - 1 items, and iteration procedures which usually begin with the squared multiple correlations and through refactoring improve the estimates upward. As is well known, the first of these estimates
(highest correlation) has no mathematical basis for use and should be avoided. Of the other two methods mentioned above, both provide reliability estimates which exceed \( \alpha \) in all cases with actual data that have been tried. Use of the squared multiple correlation typically leads to a minor increment while the iteration procedures usually provide an \( \Omega \) which is clearly higher than \( \alpha \).

The term \( \Omega \) frequently will be less than the test-retest correlation for a composite score since the latter is a function not only of latent traits but also of stabilities in the specific sources of variance for each item. Cattell and Radecki (1961) have presented a formula similar to formula (36) which substitutes test-retest correlations for items in place of \( \beta \), and their formula estimates the test-retest correlation for composite score.

Three other issues need to be mentioned. First—generally speaking—the more items that are included in a factor analysis, the higher will be the estimated communalities, since in a large analysis a greater number of small factors will be discovered that affect just complexes or triplets of items. Such minor factors, if they are true representatives of latent traits affecting a few items, ideally should be allowed to contribute to the values of \( \beta \) because they do indicate reliable sources of item variance, and one wants this reflected in \( \Omega \) (which is a function of the \( \beta \) values). Hence, as a general rule, one wants to factor as many items together as possible. Unfortunately, the general rule is complicated by two other considerations. First, some rotation procedures (like varimax) are sensitive to the number of factors extracted so that large analyses with mechanical rotation will not always yield the desired latent trait structure that would occur in a smaller, more painstaking study. Thus, a large analysis may create the need to rotate manually or to use one of the less popular rotation procedures (like quartimax) that is less sensitive to added factors. Second, there is a danger that small factors do not really represent true latent traits but develop merely as a function of sampling fluctuations, and this danger is enhanced when one analyzes many items relative to the number of observations. So, for a given sample size (N), there is a limit on how many items should be factored simultaneously; one rule of thumb specifies this limit as no more than \( N/10 \) (Nunnally, 1967, p. 257).

Second, formulas (35) and (36) for invalidity and reliability assume that \( PP = R = I + e \), that is, that the solution used to estimate factors will reproduce the observed correlation matrix (Harman, 1967, p. 63). None of the methods of estimating the \( \beta_a \) will ever exactly satisfy this assumption (Harman, 1967, p. 23), and, to the degree that this assumption is not met, results from formulas (35) and (36) will be
in error. However, when all major factors are extracted, most factoring methods will reasonably reproduce the correlation matrix, allowing that some residuals are to be expected because of sampling variations.

Finally, the formulas presented here are defined in terms of population statistics, and when these formulas are applied to sample statistics, the resulting indices would be subject to some sampling error. Since the sampling distributions of the indices are not known, the use of large samples is suggested.

REFERENCES

AMERICAN PSYCHOLOGICAL ASSOCIATION


CROSBY, J. L. 1961 "Unified models and the internal structure of tests." Psychological Science 41: 197-204.


LORD, F. M. AND NOVICE, M. R.
1968 Statistical Theory of Mental Test Scores. Reading: Addison-Wesley.
NOVICE, M. R. AND LEWIS, C.
SHERMAN, C. C.
WALKER, H. M. AND LAIT, S.
WRIGHT, S.