Guidelines for Reporting Results of Common Statistical Tests

Reliability

Reliability of a psychometric test (questionnaire, survey, etc.) is similar to the concept of precision when dealing with mechanical instrumentation, e.g., a device that measures blood sugar.

Internal Consistency Reliability

This is one form of reliability, the consistency of items within a measure. It should be reported for all measures used in a study, since it is specific to each sample.

While most authors tend to report only Cronbach’s alpha, there are other measures that may be more reasonable in various situations. Cronbach’s $\alpha$ tends to underestimate the reliability of a scale unless the condition of tau equivalence holds. If a measure contains measurement error correlations, $\alpha$ can either underestimate or overestimate reliability, depending on the underlying measurement parameters (Raykov, 2001).

Two other measures are Revelle's $\beta$ and McDonald's Omega ($\omega_h$). In unidimensional and multidimensional measures with unequal factor loadings, $\beta < \alpha < \omega_h$. Alpha can be used as an estimate of reliability with $\beta$ and $\omega_h$ representing lower and upper bounds.

See: Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005) Cronbach’s $\alpha$, Revelle's $\beta$, and McDonald’s $\omega_h$: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70, 123-133.

Another alternative to Cronbach’s $\alpha$ is to estimate scale reliability directly in the context of a CFA measurement model. The resulting reliability estimate is Raykov’s (2001) rho coefficient, which employs CFA estimates of factor loadings, error variances, and measurement error covariance to calculate factor reliabilities.

Inter-rater Reliability

Inter-rater reliability is a measure of how consistent different raters are in using a scoring or rating procedure. This should be reported for any measure that is not self-rated by the participants.

Test-retest Reliability

Test-retest reliability is a measure of how consistently a measure functions over time. This must always be reported for all measures used in a study, particularly if that study is longitudinal. The time frame for retest should be set with respect to the specific experimental design with careful consideration of whether a state or trait is being measured. States may not be expected to have much consistency over longer periods of time, e.g., moods. Traits should be stable over long periods of time, e.g., personality.
Validity

There are many different types of validity. It is important to understand the differences. The most basic types are covered here but there are others.

Content and Construct Validity

*Construct validity* means simply that a particular test actually measures the construct it is supposed to measure, e.g., a depression test measures depressive symptoms in a comprehensive and quantitative manner. This involves several aspects of test-building. Most basic is that the items of a test tap the entire content of the construct according to a formal theory; this can be considered face validity (also called *content validity*). Next, the theoretical relationships among items, or groupings of items, should be supported in a statistical sense, e.g., in a factor analysis (sometimes referred to as factorial validity).

Convergent and Divergent Validity

In its most basic form, convergent validity is shown by using a “gold standard” measure to compare to the new test: this is *criterion validity*.

The formal theory that is the basis for the test also provides information about the relationship of the construct to other measureable variables, e.g., demographics or individual traits. Standardized measures of such variables can be compared to the new test with *a priori* hypotheses about the direction and magnitude of the resultant correlations. These correlations can be significant and positive or negative (convergent validity) or non-significant (divergent validity).

Ecological and Diagnostic Validity

Ecological validity refers to the use of a test in naturalistic settings. Although a test may be developed in a laboratory setting or using an experimental manipulation of a key variable, one must be able to show that the test functions in everyday situations and in general populations. For example, a mood scale might be developed in a laboratory with college students by manipulating moods with carefully chosen film clips; but ideally, one would want to show that the test could accurately measure moods in any common situation among any demographic population.

Diagnostic validity is a more formal and strict form of ecological validity. Test used for screening or diagnosis must first be shown applicable in the populations of interest. Then, one must develop norms for the scores by comparison with a gold standard (e.g., for depression, scale scores are compared with a clinical diagnosis). *Sensitivity* and *specificity* requirements are used to develop cutting scores for diagnosis.
Effect Size and Power

According to new statistical reporting standards, it is imperative to include in a report of an empirical study (a) estimates of power based on previous literature on the topic, along with an estimate of the number of participants needed to detect the effect of interest, (b) for each statistical result, the statistic value, p value, and effect size.

There are many web sites offering advice on calculations of effect size. Some of these are:

- Sample size calculators
- Power calculators
- Calculators for Cohen's $d$ and $r$
- Effect size equations
- SPSS wiki: effect size, power, and sample size

Data Clean-up and Check of Assumptions for Analysis

One must always inspect data files for accuracy and suitability for analyses. These checks and results of any transformations should be reported in Results before any analyses. Generally this includes describing:

- What was done about missing data and outliers;
- Were normality (skewness and kurtosis), linearity, and homoscedasticity ensured or data transformed;
- Was there multicollinearity or singularity.
- Additional checks should be done for analysis of variance or covariance. These include homogeneity of regression and reliability of covariates. Sample sizes should be near equal in many designs. Choice of covariates should be cautious, since too many can reduce statistical power tremendously.

Zero-order Correlations

It is a good idea to provide correlations among your measures when you present descriptive statistics. A zero-order correlation matrix must be provided when any type of regression is run. A table is the best way to do this:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>Variable 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Place internal consistency reliabilities on the diagonal in boldface font. Correlations of .20 or so are meaningless without a strong theoretical context. Any randomly chosen psychological variables will be correlated at approximately that level, as demonstrated by Paul Meehl. Be careful about interpreting those. Also, correlations should be accompanied by 95% confidence intervals, e.g., $r\ (L, U)$.
The t test

Required reporting:
- The value of $t$
- The degrees of freedom ($df$)
- The $p$ level; exact level should be given, rather than $p < .05$. Use exponential notation if needed, e.g., write $p = 8.2\times10^{-3}$; or simply report $p < .001$.
- The effect size (Cohen's $d = \text{standardized mean difference}$)

If the results of the $t$-test are reported in the text, it can be written as:

"$t_{27} = 3.27, p = .003, d = 0.52$"

If the results of a number of $t$ tests are given in a table (here, shown with group means), it should look like:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group 1</th>
<th>Group 2</th>
<th>$t$</th>
<th>$df$</th>
<th>$p$</th>
<th>Cohen's $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yr.</td>
<td>23.2</td>
<td>25.4</td>
<td>3.27</td>
<td>27</td>
<td>.003</td>
<td>0.52</td>
</tr>
<tr>
<td>Height, m</td>
<td>1.70</td>
<td>1.71</td>
<td>1.29</td>
<td>32</td>
<td>ns</td>
<td>0.13</td>
</tr>
</tbody>
</table>

The $\chi^2$ test

Requirements
- The value of $\chi^2$
- The degrees of freedom ($df$)
- The $p$ level (exact level should be given, rather than $p < .05$)
- The effect size ($\phi$ for $2 \times 2$ tables)

If the results of the $\chi^2$ test are reported in the text, it can be written as:

"$\chi^2_1 = 4.23, p = .04, \phi = 0.29$"

If the results of a number of $\chi^2$ tests are given in a table, it should look like:

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\chi^2$</th>
<th>$df$</th>
<th>$p$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>4.23</td>
<td>1</td>
<td>.04</td>
<td>0.29</td>
</tr>
<tr>
<td>Marital status</td>
<td>2.58</td>
<td>5</td>
<td>ns</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Guidelines for Reporting Results of Statistical Tests

http://www.amsci.com/authors/author-resources/
Analysis of Variance (ANOVA)

Required reporting:

- The value of $F$
- The numerator and denominator degrees of freedom ($df$)
- The $p$ level (exact level should be given, rather than $p<.05$)
- If the $p$ level is < .05 and there are more than two groups, the results of a post hoc test.
- The effect size ($\eta^2$ or $\omega^2$)

If the results of the ANOVA are reported in the text, it can be written as:

"$F_{2,27}= 4.01, p = .03, \eta^2 = 0.096$. The Tukey post hoc test showed Group 1 scored significantly higher than Group 3."

If the results of an ANOVA with two or more main effects are given in a table, it should look like this:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>3</td>
<td>9.29</td>
<td>11.99</td>
<td>&lt;.001</td>
<td>0.22</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>18.23</td>
<td>23.52</td>
<td>&lt;.001</td>
<td>0.14</td>
</tr>
<tr>
<td>Group x Gender</td>
<td>3</td>
<td>19.49</td>
<td>25.15</td>
<td>&lt;.001</td>
<td>0.42</td>
</tr>
<tr>
<td>Error</td>
<td>32</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Multiple Regression

A table for a model summary should be arranged like this:

<table>
<thead>
<tr>
<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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\Delta R^2$</td>
<td>$\Delta F$</td>
<td>$df_1$</td>
<td>$df_2$</td>
<td>$p$</td>
</tr>
<tr>
<td>1&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Predictors: (Constant), Variable 1.  <sup>b</sup>Predictors: (Constant), Variable 1, Variable 2.

A table for coefficients should be arranged like this:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized coefficients</th>
<th>Standardized coefficients</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>$\beta$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>(Constant)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>(Constant)</td>
<td>Variable 1 Variable 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Guidelines for Reporting Results of Statistical Tests  
[http://www.amsci.com/authors/author-resources/](http://www.amsci.com/authors/author-resources/)
Heirarchical Regression

Below is a layout with the example assuming three steps and two variables entered at each step. The reader can easily see how the Bs and betas do or don’t change as new variables are added. (Stepwise regression: use only if the aim is to arrive at a parsimonious predictive equation, and you don’t care why the variables are or are not in the equation.)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>SE</td>
<td>β</td>
</tr>
<tr>
<td>A</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>B</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>C</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>D</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>E</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>F</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>R²</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>AdjR²</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>SE</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>F(df, df)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Factor Analysis

If you do a factor analysis with a subject-to-variable ratio less than 10, you’ll get an output (indeed, you’ll get one even if there are fewer subjects than variables, although the computer really should have had a major infarct). But, there will be many problems: variables will likely load on the wrong factor; the eigenvalues will be wrong; the factor structure may be wrong; and you may end up with Heywood cases, in which the factor loading exceeds 1.0 (Costello & Osborne, 2005). Remember, factor analysis is a large-sample procedure!

Required reporting:

If you run a factor analysis using any of the statistical packages, you’ll get reams of output. The issue is then what should you report when you write up your results.

- The sample size and subject-to-variable ratio.
- The method of factor extraction (e.g., principal axis, principal components analysis).
- The criteria used to determine the number of factors to retain: eigenvalues, scree test, coherence of item meanings within factors. Report how you used these criteria to make decisions, particularly about which factors and items to retain.

Guidelines for Reporting Results of Statistical Tests  
http://www.amsci.com/authors/author-resources/
• The method of factor rotation, indicating first whether it was orthogonal or oblique, and then the specific type. Justify your choice—don’t use orthogonal rotation if the factors are likely to be correlated.
• The entire factor loading matrix. Some people don’t report loadings that are below some minimum value in order to make the matrix more readable. But this makes it harder for the reader to see if there is any factorial complexity. Provide all the loadings, indicating in boldface which ones are retained for each factor.
• If an oblique rotation is used, you should also report the factor structure matrix and the correlation matrix among the factors.
• The eigenvalues of the factors after rotation and the total amount of variance accounted for.
• The communalities of all variables.

**Structural Equation Modeling & Confirmatory Factor Analysis**

Programs for SEM and CFA produce reams and reams of output. This in turn has resulted in reams and reams of guidelines for reporting the results. Dr. David Streiner has summarized this set of guidelines from multiple sources.

• The most helpful thing to report is the diagram of the model, with the standardized regression weights and correlations near the arrows. You should also indicate which ones are statistically significant.
• In the text, there should be a theoretical or research-based rationale for every arrow in the diagram, both directional (one arrow head) and non-directional (the curved arrows with two heads).
• There are many programs for SEM and CFA, and they differ with regard to their methods. This means that, unlike with techniques such as ANOVA or EFA, different programs can produce different results. Consequently, you must say which program you used.
• You should indicate what kind of matrix was analyzed (covariances, covariance with means, correlation, etc.). If you don’t know, you shouldn’t be using these techniques.
• Never forget that CFA and SEM are large-sample techniques. You need to report not only the sample size but also the ratio of subjects to parameters (this implies that you know how many parameters you’re estimating).
• How you treated outliers and missing data – did you drop those cases, impute values, or use some technique that can handle these anomalies?
• What method did you use for parameter estimation – maximum likelihood, generalized least squares, weighted least squares, etc. – and why?
• The model fit. As we said, there are many indices of this, but we would recommend as a minimum: (a) and its df; (b) the Tucker-Lewis Index (TLI, a.k.a. the NNFI); (c) the RMSEA; (d) SRMSR; and (e) the NFI.
• Indicate if you’ve changed the model on the basis of the modification indices. If you have, state what was changed and the theoretical justification for it, as well as how much the modifications improved the fit.
Resources For Specific Statistics Questions

Articles by David Streiner

Experimental design and reporting
- Sample size and power
- Graphing techniques
- Confounding variables
- Writing a Results section
- Qualitative vs quantitative methods
- Regression toward the mean

Basic statistical concepts
- Central tendency
- Standard deviation vs standard error
- Agreement and reliability
- Dichotomizing continuous data
- Multivariate statistics

Advanced statistical techniques
- Checklist for rating scales
- Item response theory
- Factor analysis
- Stepwise multiple regression
- Estimate of risk
- Survival analysis
- Path analysis
- Structural equation modeling

Also of interest:

Andrew Vickers on using parametric versus non-parametric statistics

SAS statistics by example (free e-book download)

Martin Eppler’s project for teaching teachers and team leaders about how to use graphical techniques for communicating data.