An introduction to Psychometric Theory

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

April, 2020
Outline of Day 1/part 1

What is psychometrics?
  Goals and Requirements

Conceptual overview
  Theory: the organization of Observed and Latent variables

Overview
  A latent variable approach to measurement
    Types of validity; What are we measuring
    Structural Equation Models
    Individual differences

Two disciplines of scientific psychology
  Two cultures
  Two tribes within the scientific culture

Theory testing
  Persistence of theories
  The process of theory testing

What is R? Where did it come from, why use it?
  Installing R on your computer and adding packages
  Installing and using packages
What is psychometrics?

In physical science a first essential step in the direction of learning any subject is to find principles of numerical reckoning and methods for practicably measuring some quality connected with it. I often say that when you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the stage of science, whatever the matter may be. (?)

What is psychometrics?

The character which shapes our conduct is a definite and durable ‘something’, and therefore . . . it is reasonable to attempt to measure it. (?

The history of science is the history of measurement” (J. M. Cattell, 1893)

Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality (E.L. Thorndike, 1918). See also ?.
What is psychometrics?

We hardly recognize a subject as scientific if measurement is not one of its tools (Boring, 1929)

There is yet another [method] so vital that, if lacking it, any study is thought ... not be scientific in the full sense of the word. This further an crucial method is that of measurement. (Spearman, 1937)

One’s knowledge of science begins when he can measure what he is speaking about and express in numbers (Eysenck, 1973)

Psychometrics: the assigning of numbers to observed psychological phenomena and to unobserved concepts. Evaluation of the fit of theoretical models to empirical data.
Goals

1. To acquire the fundamental vocabulary and logic of psychometric theory.
2. To develop your capacity for critical judgment of the adequacy of measures purported to assess psychological constructs.
3. To acquaint you with some of the relevant literature in personality assessment, psychometric theory and practice, and methods of observing and measuring affect, behavior, cognition and motivation.
4. To instill an appreciation of and an interest in the principles and methods of psychometric theory.
5. This course is not designed to make you into an accomplished psychometrist (one who gives tests) nor is it designed to make you a skilled psychometrician (one who constructs tests).
6. It will give you limited experience with psychometric computer programs (although all of the examples will use R, it is not necessary to learn R, but it sure helps.).
Requirements

1. Asking questions!
2. Objective Midterm exam
3. Objective Final exam
4. Final paper applying principles from the course to a problem of interest to you.
5. Sporadic applied homework and data sets
Texts (required and recommended supplements)

   http://www.personality-project.org/r/book/[An introduction to psychometric theory with applications in R]. Springer
   Available on line and linked throughout the syllabus.

2. ? Useful reading on SEM (suggested to read)

3. ? An old classic but dreadfully out of date.
Psychometric Theory: A conceptual Syllabus

Error $\delta_1$ $\delta_2$ $\delta_3$ $\delta_4$ $\delta_5$ $\delta_6$ $\delta_7$ $\delta_8$ $\delta_9$ $X_1$ $X_2$ $X_3$ $X_4$ $X_5$ $X_6$ $X_7$ $X_8$ $X_9$ $\chi_1$ $\chi_2$ $\chi_3$ $\xi_1$ $\xi_2$ $Y_1$ $Y_2$ $Y_3$ $Y_4$ $Y_5$ $Y_6$ $Y_7$ $Y_8$ Error $\epsilon_1$ $\epsilon_2$ $\epsilon_3$ $\epsilon_4$ $\epsilon_5$ $\epsilon_6$ $\epsilon_7$ $\epsilon_8$
Observed Variables

Error \quad X \quad \text{Latent X} \quad \text{Latent Y} \quad Y \quad Error

\delta_1 \quad X1 \quad \chi_1 \quad \xi_1 \quad Y1 \quad \epsilon_1
\delta_2 \quad X2
\delta_3 \quad X3
\delta_4 \quad X4
\delta_5 \quad X5
\delta_6 \quad X6
\delta_7 \quad X7
\delta_8 \quad X8
\delta_9 \quad X9

\chi_2 \quad \xi_2 \quad Y2
\chi_3 \quad Y3
\chi_2 \quad Y4
\chi_3 \quad Y5
\chi_3 \quad Y6
\chi_3 \quad Y7
\chi_3 \quad Y8
Latent Variables

X

\[ \delta_1 \rightarrow X_1 \]
\[ \delta_2 \rightarrow X_2 \]
\[ \delta_3 \rightarrow X_3 \]
\[ \delta_4 \rightarrow X_4 \]
\[ \delta_5 \rightarrow X_5 \]
\[ \delta_6 \rightarrow X_6 \]
\[ \delta_7 \rightarrow X_7 \]
\[ \delta_8 \rightarrow X_8 \]
\[ \delta_9 \rightarrow X_9 \]

Latent X

\[ \chi_1 \]
\[ \chi_2 \]
\[ \chi_3 \]

Latent Y

\[ \xi_1 \]
\[ \xi_2 \]

Error

\[ \delta_1 \rightarrow \xi_1 \]
\[ \delta_2 \rightarrow \xi_1 \]
\[ \delta_3 \rightarrow \xi_1 \]
\[ \delta_4 \rightarrow \xi_1 \]
\[ \delta_5 \rightarrow \xi_1 \]
\[ \delta_6 \rightarrow \xi_1 \]
\[ \delta_7 \rightarrow \xi_1 \]
\[ \delta_8 \rightarrow \xi_1 \]
\[ \delta_9 \rightarrow \xi_1 \]

\[ \epsilon_1 \rightarrow Y_1 \]
\[ \epsilon_2 \rightarrow Y_2 \]
\[ \epsilon_3 \rightarrow Y_3 \]
\[ \epsilon_4 \rightarrow Y_4 \]
\[ \epsilon_5 \rightarrow Y_5 \]
\[ \epsilon_6 \rightarrow Y_6 \]
\[ \epsilon_7 \rightarrow Y_7 \]
\[ \epsilon_8 \rightarrow Y_8 \]
Theory

Error \ X \ Latent X \ Latent Y \ Y \ Error

\[ \delta_1 \rightarrow X_1 \rightarrow \chi_1 \rightarrow \xi_1 \]
\[ \delta_2 \rightarrow X_2 \]
\[ \delta_3 \rightarrow X_3 \]
\[ \delta_4 \rightarrow X_4 \]
\[ \delta_5 \rightarrow X_5 \rightarrow \chi_2 \rightarrow \xi_2 \]
\[ \delta_6 \rightarrow X_6 \]
\[ \delta_7 \rightarrow X_7 \]
\[ \delta_8 \rightarrow X_8 \]
\[ \delta_9 \rightarrow X_9 \]

\[ \chi_3 \rightarrow \xi_2 \]
\[ \xi_1 \rightarrow Y_1 \]
\[ \xi_1 \rightarrow Y_2 \]
\[ \xi_1 \rightarrow Y_3 \]
\[ \xi_1 \rightarrow Y_4 \]
\[ \xi_2 \rightarrow Y_5 \]
\[ \xi_2 \rightarrow Y_6 \]
\[ \xi_2 \rightarrow Y_7 \]
\[ \xi_2 \rightarrow Y_8 \]

\[ \epsilon_1 \rightarrow Y_1 \]
\[ \epsilon_2 \rightarrow Y_2 \]
\[ \epsilon_3 \rightarrow Y_3 \]
\[ \epsilon_4 \rightarrow Y_4 \]
\[ \epsilon_5 \rightarrow Y_5 \]
\[ \epsilon_6 \rightarrow Y_6 \]
\[ \epsilon_7 \rightarrow Y_7 \]
\[ \epsilon_8 \rightarrow Y_8 \]
Psychometric Theory: A conceptual Syllabus

Error $\delta_1$ $\delta_2$ $\delta_3$ $\delta_4$ $\delta_5$ $\delta_6$ $\delta_7$ $\delta_8$ $\delta_9$

Latent X $\chi_1$

Error $\epsilon_1$ $\epsilon_2$ $\epsilon_3$ $\epsilon_4$ $\epsilon_5$ $\epsilon_6$ $\epsilon_7$ $\epsilon_8$

Latent Y $\xi_1$ $\xi_2$

$Y_1$ $Y_2$ $Y_3$ $Y_4$ $Y_5$ $Y_6$ $Y_7$ $Y_8$

$X_1$ $X_2$ $X_3$ $X_4$ $X_5$ $X_6$ $X_7$ $X_8$ $X_9$
A theory of data and fundamentals of scaling
Correlation, Regression, Partial Correlation, Multiple Regression

Error \[ \delta_1 \]

\begin{align*}
X1 & \rightarrow \beta_{y,x} \rightarrow Y1 \\
X2 & \rightarrow \beta_{x,y} \rightarrow Y2 \\
X3 & \rightarrow r_{xy} \rightarrow Y3 \\
X4 & \rightarrow r_{x_4y_4.x_5} \rightarrow Y4 \\
X5 & \rightarrow R_{y.x_6x_7x_8} \rightarrow Y5 \\
X6 & \rightarrow Y6 \\
X7 & \rightarrow Y7 \\
X8 & \rightarrow Y8 \\
X9 & \rightarrow Error
\end{align*}

Error \[ \epsilon_1 \] \[ \epsilon_2 \] \[ \epsilon_3 \] \[ \epsilon_4 \] \[ \epsilon_5 \] \[ \epsilon_6 \] \[ \epsilon_7 \] \[ \epsilon_8 \]
Measurement: A latent variable approach.

Error  X  Latent X

Latent Y  Y  Error

\[ \begin{align*}
\delta_1 & \rightarrow X_1 \\
\delta_2 & \rightarrow X_2 \\
\delta_3 & \rightarrow X_3 \\
\delta_4 & \rightarrow X_4 \\
\delta_5 & \rightarrow X_5 \\
\delta_6 & \rightarrow X_6 \\
\delta_7 & \rightarrow X_7 \\
\delta_8 & \rightarrow X_8 \\
\delta_9 & \rightarrow X_9 \\
\chi_1 & \rightarrow \xi_1 \\
\chi_2 & \rightarrow \xi_2 \\
\chi_3 & \\
\end{align*} \]
Reliability: How well does a test reflect one latent trait?

Error \( \delta_1 \) \( \delta_2 \) \( \delta_3 \) \( \delta_4 \) \( \delta_5 \) \( \delta_6 \) \( \delta_7 \) \( \delta_8 \) \( \delta_9 \)

\( X_1 \) \( X_2 \) \( X_3 \) \( X_4 \) \( X_5 \) \( X_6 \) \( X_7 \) \( X_8 \) \( X_9 \)

Latent \( \chi_1 \) \( \chi_2 \) \( \chi_3 \)

Latent \( \xi_1 \) \( \xi_2 \) \( \xi_3 \)

\( Y_1 \) \( Y_2 \) \( Y_3 \) \( Y_4 \) \( Y_5 \) \( Y_6 \) \( Y_7 \) \( Y_8 \)

Error \( \epsilon_1 \) \( \epsilon_2 \) \( \epsilon_3 \) \( \epsilon_4 \) \( \epsilon_5 \) \( \epsilon_6 \) \( \epsilon_7 \) \( \epsilon_8 \)
Face, Concurrent, Predictive, Construct

Error \[ \delta_1 \quad \delta_2 \quad \delta_3 \quad \delta_4 \quad \delta_5 \quad \delta_6 \quad \delta_7 \quad \delta_8 \quad \delta_9 \]

X \[ X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9 \]

Latent X \[ \chi_2 \]

Concurrent \[ \chi_2 \quad Y_1 \quad \epsilon_1 \]

Predictive \[ \chi_2 \quad Y_2 \quad \epsilon_2 \]

Construct \[ \chi_2 \quad Y_3 \quad \epsilon_3 \]

\[ \chi_2 \quad Y_4 \quad \epsilon_4 \]

\[ \chi_2 \quad Y_5 \quad \epsilon_5 \]

\[ \chi_2 \quad Y_6 \quad \epsilon_6 \]

\[ \chi_2 \quad Y_7 \quad \epsilon_7 \]

\[ \chi_2 \quad Y_8 \quad \epsilon_8 \]
Psychometric Theory: Data, Measurement, Theory

Error \( \delta_1 \) \( \delta_2 \) \( \delta_3 \) \( \delta_4 \) \( \delta_5 \) \( \delta_6 \) \( \delta_7 \) \( \delta_8 \) \( \delta_9 \) \( \chi_1 \) \( \chi_2 \) \( \chi_3 \) \( \xi_1 \) \( \xi_2 \) \( Y_1 \) \( Y_2 \) \( Y_3 \) \( Y_4 \) \( Y_5 \) \( Y_6 \) \( Y_7 \) \( Y_8 \) \( \epsilon_1 \) \( \epsilon_2 \) \( \epsilon_3 \) \( \epsilon_4 \) \( \epsilon_5 \) \( \epsilon_6 \) \( \epsilon_7 \) \( \epsilon_8 \)

\[ \begin{align*}
\delta_1 & \rightarrow X_1 \\
\delta_2 & \rightarrow X_2 \\
\delta_3 & \rightarrow X_3 \\
\delta_4 & \rightarrow X_4 \\
\delta_5 & \rightarrow X_5 \\
\delta_6 & \rightarrow X_6 \\
\delta_7 & \rightarrow X_7 \\
\delta_8 & \rightarrow X_8 \\
\delta_9 & \rightarrow X_9 \\
\chi_1 & \\
\chi_2 & \\
\chi_3 & \\
\xi_1 & \rightarrow Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8 \\
\xi_2 & \\
\epsilon_1 & \\
\epsilon_2 & \\
\epsilon_3 & \\
\epsilon_4 & \\
\epsilon_5 & \\
\epsilon_6 & \\
\epsilon_7 & \\
\epsilon_8 & \\
\end{align*} \]
Psychometric Theory: Data, Measurement, Theory

Error → X → Latent X → Avoidance → Latent Y → Y → Error

Avoidance

Approach

Cog

Ability

\[ \delta_1 \rightarrow X_1 \]
\[ \delta_2 \rightarrow X_2 \]
\[ \delta_3 \rightarrow X_3 \]
\[ \delta_4 \rightarrow X_4 \]
\[ \delta_5 \rightarrow X_5 \]
\[ \delta_6 \rightarrow X_6 \]
\[ \delta_7 \rightarrow X_7 \]
\[ \delta_8 \rightarrow X_8 \]
\[ \delta_9 \rightarrow X_9 \]
Overview

1. Overview: An introduction to Psychometric Theory
   • What is Psychometrics?
   • What is R?


3. Weeks 2-3: More than you ever wanted to know about correlation.
   • A review (?) of linear algebra.

4. Weeks 4-5 Dimension reduction through factor analysis, principal components analyze and cluster analysis.

5. Week 6-7 Classical Test Theory and Item Response Theory.

6. Week 8-9 Structural Equation Modeling and applied scale construction.
The process of science

Prologue: two broad themes to be discussed and interwoven

1. The two disciplines of scientific psychology
   1.1 Two broad cultures of intellectual activity (?)
   1.2 Two broad cultures of psychology (?)
   1.3 Two disciplines within scientific psychology (??) and (???).

2. The process of theory construction and validation
   2.1 Science from hunch to law (??)
   2.2 Good theories as alive and generative: the example of theories of Extraversion.

I will emphasize the power of integrating psychometric and experimental techniques in a programmatic study of personality and individual differences. This is mainly a shameless advertisement for interactive personality research.
C.P. ? considered two cultures of intellectual inquiry:

“I believe the intellectual life of the whole of western society is increasingly being split into two polar groups.” ..

“I felt I was moving among two groups—comparable in intelligence, identical in race, not grossly different in social origin, earning about the same incomes, who had almost ceased to communicate at all, who in intellectual, moral and psychological climate had so little in common ... one might have crossed an ocean.”
Kimble and the two cultures of psychology

Just as Snow considered the scientific versus humanistic cultures of English and American society, so did Kimble consider two cultures of psychology: the scientific and the humanistic.

“The remaining points of disagreement involve the items asking about most important values (scientific vs. human), source of basic knowledge (objectivism vs. intuitionism), and generality of laws (nomothetic vs. idiographic).
Two competing tribes/paradigms within scientific psychology

1. But even within the culture of scientific psychology, we have two competing tribes who differ in their basic paradigmatic view of how to do science: the correlational vs. experimental paradigms discussed by ?? and ????. Both pleaded for an integration of the two tribes. Neither was overly successful.

2. Others who have tried to reconcile these differences included ?, and ?.

3. ? reported that this dichotomy still continues.

4. We need to go beyond this dichotomy by realizing that theory development and theory testing require a mixture of the inductive power of correlations with the deductive power of experimental techniques.

5. For we as individual differences psychologists are most able to unify the two disciplines.
The conventional dichotomy of research paradigms in psychology ala ?? and ???

Correlational

1. Influential founders
   1.1 ?
   1.2 ?
   1.3 ?

2. Measurement of variances and covariances
   2.1 bivariate r, φ, YuleQ
   2.2 multivariate R, factor analysis, principal components
   2.3 General Linear Model and its extension to multi-level modeling

3. Addresses threats to validity by statistical “control”

Experimental

1. Influential founders
   1.1 ?
   1.2 Gossett (?)
   1.3 ?

2. Measurement of central tendencies
   2.1 bivariate t and F
   2.2 multivariate MANOVA

   2.3 General Linear Model and its extension to multi-level modeling

3. Addresses threats to validity by randomization
Two disciplines: two viewpoints

Table: The naive perspective from both sides—the other side is easy, why don’t they just do it right? Our variables are complicated, well articulated, theirs are simple, just use any one.

<table>
<thead>
<tr>
<th>Individual Differences</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality Ability</td>
<td>Task Performance</td>
</tr>
</tbody>
</table>
The experimentalist’s challenge: what to measure

Measures
1. Giant 3
   - EPI
   - EPQ

2. Big 5
   - NEO-PI-R
   - IPIP B5
   - IPIP NEO
   - BFI
   - TIPI

3. Beyond the Big 5
   - HEXACO
   - IPIP HEXACO
   - BFAS
   - SAPA 3-6-12-27
   - ICAR-IQ
   - ...

Constructs
1. Extraversion
   - but which one? Costa vs. Goldberg

2. Neuroticism

3. Agreeableness

4. Conscientiousness

5. Openness-Intellect
   - but is it openness or is it intellect?

6. Honesty/Humility

7. Impulsivity

8. Sociability

9. Trust

10. ...
The challenge for individual difference researchers: what constructs to measure

Memory

1. Working memory
2. Iconic memory
3. Short Term memory
4. Long Term memory
5. Semantic memory
6. Episodic memory
7. Procedural memory
8. Autobiographical memory
9. False memory
10. Recall
11. Recognition

Attention

1. Sustained Attention
2. Allocation of Attention
3. Capturing Attention
4. Breadth of Attention
5. Local/Global Attention
6. Paying Attention

System I or System II

1. Fast, automatic
2. Slow, controlled, but lazy
The experimentalist’s challenge: how to analyze, what to report

**Analysis**

1. Dimension Reduction
   - Principal Components
   - EFA
   - CFA

2. Structure
   - Path Analysis
   - SEM
   - Latent Growth Curves

3. Reliability analysis
   - Internal Consistency
   - Alternate Form
   - Test-Retest

4. Item Response Theory

**Statistics**

1. Measures of association
   - Pearson $r$, Spearman $\rho$
   - $\phi$ or Yule $Q$
   - $r_{tetrachoric}$, $r_{polychoric}$

2. Goodness of fit
   - $\chi^2$ or $\chi^2$ difference
   - RMSEA or RMSR
   - Tucker-Lewis
   - BIC or AIC

3. Reliability
   - $\alpha$
   - $\beta$
   - $\omega_h$
   - $\omega_t$
The challenge for individual difference researchers: which paradigm to use

Memory
1. Reaction time
   • Sternberg Memory Scanning
   • Ratcliff choice
   • Jacoby identification
2. Accuracy
3. Serial anticipation
4. Free recall
5. Cued recognition

Attention
1. Posner letter search
2. Erickson flanker task
3. Vigilance
4. dot probe
5. emotional “Stroop”
6. Eye tracking
7. Reaction Time
The extra subtleties of design

Personality
1. Item wording
2. Response alternatives
3. Appropriate sample size
4. Subject selection
   • restriction of range
5. Generalization of subject characteristics

Experimental
1. Number of practice trials
2. Inter Stimulus Interval
3. Stimulus Onset Asynchrony
4. Type of randomization/counterbalancing
   • block randomization
   • complete randomization
   • counterbalancing
5. Data trimming procedures
6. Power/p-hacking
Theory testing is hard work

1. Confirmatory bias
2. Theory induced blindness
3. Seductive power of hindsight
4. Illusion of control
   • Under appreciation of chance
5. See *Thinking, Fast and Slow*?
Scientific progress and levels of theory

1. Hunch
   - observations
   - deduction

2. Hypothesis
   - hypothesis development
   - hypothesis verification

3. Theory
   - Weak theory – confirmation studies
   - Strong theory – disconfirmation studies

4. Law
Eysenck, Lakatos, Popper and Kuhn

followed in suggesting that disconfirmation studies did not lead to theory rejection until a better theory was supplied.

“Purely negative, destructive criticism, like ‘refutation’ or demonstration of an inconsistency does not eliminate a programme. Criticism of a programme is a long and often frustrating process and one must treat budding programmes leniently. One can, of course, undermine a research-programme but only with dogged patience. It is usually only constructive criticism which, with the help of rival research programmes can achieve major successes; but even so, dramatic, spectacular results become visible only with hindsight and rational reconstruction.” (?, p 183)
Eysenck’s theory as an adaptive and changing theory of personality

thought that the building of paradigmatic personality research required critical analysis of theory and welcomed the publications of some of his strongest critics (e.g., ?).

“the existence of anomalies should be no bar to the acceptance of the paradigm; the existence of such anomalies should merely act as a spur for the puzzle-solving capacities of ordinary science.”

Indeed, in his presidential address to the International Society for the Study of Individual Differences ? spent much of the time discussing Gray’s criticisms and then cheerfully announced that Gray was going to replace him at the Maudsley!
Use R
R: Statistics for all of us

1. What is it?
2. Why use it?
3. Common (mis)perceptions of R
4. Examples for psychologists
   - graphical displays
   - basic statistics
   - advanced statistics
   - Although programming is easy in R, that is beyond the scope of this course.
R: What is it?

1. R: An international collaboration
2. R: The open source - public domain version of S+
3. R: Written by statistician (and all of us) for statisticians (and the rest of us)
4. R: Not just a statistics system, also an extensible language.
   • This means that as new statistics are developed they tend to appear in R far sooner than elsewhere.
   • R facilitates asking questions that have not already been asked.
5. R: encourages publications of "Reproducible Research"
   • integrate data, code, text into one document
   • Sweave and knitr
6. Many journals and chapters include R code appendices to allow for open science.
Statistical Programs for Psychologists

• General purpose programs
  • R
  • S+
  • SAS
  • SPSS
  • STATA
  • Systat

• Specialized programs
  • Mx
  • EQS
  • AMOS
  • LISREL
  • MPlus
  • Your favorite program
Statistical Programs for Psychologists

• General purpose programs
  • R
  • $+$
  • $A$
  • $P$
  • $TATA$
  • $y$stat

• Specialized programs
  • Mx (OpenMx is part of R)
  • EQ$
  • AMO$
  • LI$REL
  • MPlu$
  • Your favorite program
R: A way of thinking

• “R is the lingua franca of statistical research. Work in all other languages should be discouraged.” (Jan de Leeuw, 2003)

• “This is R. There is no if. Only how.” (Simon ‘Yoda’ Blomberg, 2005)

• “Overall, SAS is about 11 years behind R and S-Plus in statistical capabilities (last year it was about 10 years behind) in my estimation.” (Frank Harrell, 2003)

• ”I quit using SAS in 1991 because my productivity jumped at least 20% within one month of using S-Plus.” (Frank Harrell, 2003)

Taken from the R.-fortunes (selections from the R.-help list serve)
More fortunes

“You must realize that R is written by experts in statistics and statistical computing who, despite popular opinion, do not believe that everything in SAS and SPSS is worth copying. Some things done in such packages, which trace their roots back to the days of punched cards and magnetic tape when fitting a single linear model may take several days because your first 5 attempts failed due to syntax errors in the JCL or the SAS code, still reflect the approach of “give me every possible statistic that could be calculated from this model, whether or not it makes sense”. The approach taken in R is different. The underlying assumption is that the useR is thinking about the analysis while doing it. ” (Douglas Bates, 2007)
R is open source, how can you trust it?

• Q: “When you use it [R], since it is written by so many authors, how do you know that the results are trustable?”

• A: “The R engine [...] is pretty well uniformly excellent code but you have to take my word for that. Actually, you don’t. The whole engine is open source so, if you wish, you can check every line of it. If people were out to push dodgy software, this is not the way they’d go about it.” (Bill Venables, 2004)

• “It’s interesting that SAS Institute feels that non-peer-reviewed software with hidden implementations of analytic methods that cannot be reproduced by others should be trusted when building aircraft engines.” – Frank Harrell (in response to the statement of the SAS director of technology product marketing: "We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet.”) R-help (January 2009)
What is R?: Technically

- R is an open source implementation of S (S-Plus is a commercial implementation)
- R is available under GNU Copy-left
- R1.0.0 was released on February 29, 2000
- The current version of R is 3.6.3
- R is a group project run by a core group of developers (with new releases ≈ semiannually)
- R 4.0.0 is to be released sometime in late April, 2020

(Adapted from Robert Gentleman)
**R: A brief history**

- **1991-93:** Ross Dhaka and Robert Gentleman begin work on R project for Macs at U. Auckland (S for Macs).
- **1995:** R available by ftp under the General Public License.
- **96-97:** mailing list and R core group is formed.
- **2000:** John Chambers, designer of S joins the Rcore (wins a prize for best software from ACM for S)
- **2001-2019:** Core team continues to improve base package with a new release every 6 months (now more like yearly).
- Many others contribute “packages” to supplement the functionality for particular problems.
  - 2003-04-01: 250 packages
  - 2004-10-01: 500 packages
  - 2007-04-12: 1,000 packages
  - 2009-10-04: 2,000 packages
  - 2011-05-12: 3,000 packages
  - 2012-08-27: 4,000 packages
  - 2014-05-16: 5,547 packages (on CRAN) + 824 bioinformatic packages on BioConductor
  - 2016-03-21 8,120 packages (on CRAN) + 1,104 bioinformatic packages + ?,000s on GitHub/R-Forge
  - 2017-03-26 10,327 packages (CRAN) + 1,296 on BioConductor + ?,000s on GitHub
  - 2020-04-04 15,514 packages (CRAN) + 1,823 on BioConductor + ?,000s on GitHub
Has R grown too much? Exponential growth rate continues

See also [http://r4stats.com/articles/popularity/](http://r4stats.com/articles/popularity/)
Misconception: R is hard to use

1. R doesn’t have a GUI (Graphical User Interface)
   - Partly true, many use syntax
   - Partly not true, GUls exist (e.g., R Commander, R-Studio)
   - Quasi GUls for Mac and PCs make syntax writing easier

2. R syntax is hard to use
   - Not really, unless you think an iPhone is hard to use
   - Easier to give instructions of 1-4 lines of syntax rather than pictures of what menu to pull down.
   - Keep a copy of your syntax, modify it for the next analysis.

3. R is not user friendly: A personological description of R
   - R is introverted: it will tell you what you want to know if you ask, but not if you don’t ask.
   - R is conscientious: it wants commands to be correct.
   - R is not agreeable: its error messages are at best cryptic.
   - R is stable: it does not break down under stress.
   - R is open: new ideas about statistics are easily developed.
Misconceptions: R is hard to learn

1. With a brief web based tutorial
   http://personality-project.org/r, 2nd and 3rd year undergraduates in psychological methods and personality research courses are using R for descriptive and inferential statistics and producing publication quality graphics.

2. More and more psychology departments are using it for graduate and undergraduate instruction.

3. R is easy to learn, hard to master
   - R-help newsgroup is very supportive
   - Multiple web based and pdf tutorials see (e.g., http://www.r-project.org/)
   - Short courses using R for many applications

4. Books and websites for SPSS and SAS users trying to learn R (e.g., http://oit.utk.edu/scc/RforSAS&SPSSusers.pdf by Bob Muenchen).
Ok, how do I get it: Getting started with R

1. Download from R Cran (http://cran.r-project.org/)
   - Choose appropriate operating system and download compiled R

2. Install R (current version is 3.6.3) with 4.0.0 coming this spring

3. Start R

4. Add useful packages (just need to do this once)
   - install.packages("ctv") #this downloads the task view package
   - library(ctv) #this activates the ctv package
   - install.views("Psychometrics") #among others
   - Take a 5 minute break

5. Activate the package(s) you want to use (e.g., psych)
   - library(psych) #necessary for most of this course
   - library(lavaan) #will be used for a few examples

6. Use R

7. (See detailed tutorial at https://personality-project.org/r/r.guide.html#gettingstarted)
Start up R and get ready to play (Mac version)

R Under development (unstable) (2020-02-21 r77847) -- "Unsuffered Consequences"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.70 (7782) x86_64-apple-darwin15.6.0]

[Workspace restored from /Users/WR/.RData]
[History restored from /Users/WR/.Rapp.history]

Good evening Bill.
Annotated installation guide: don’t type the 

```r
#To install all the psychometric task views:
> install.packages("ctv")
> library(ctv)
> install.views("Psychometrics")

#or just install a few packages
> install.packages("psych")
> install.packages("psychTools")
> install.packages("GPArotation")
> install.packages("mnormt")
```

Install the task view installer package. You might have to choose a “mirror” site.

- Make it active
- Install all the packages in the “Psychometrics” task view. This will take a few minutes.
- Or, just install one package (e.g., psych) as well as a few suggested packages that add functionality for factor rotation, multivariate normal distributions, etc.
The psych package is under constant development

1. The most recent “official version” is on CRAN
2. The development version is on the pmc server and may be installed from there:
   - `install.packages("psych", repos="http://personality-project.org/r", type="source")`
3. This will give you the “bleeding edge” version.
   - Bug fixes have been added.
   - New features have been added.
4. Read the news to find out what has changed.
   - `news(Version >"1.9.12", package="psych")`
Check the version number for R (should be $\geq 3.6.3$ and for psych ($\geq 1.9.12$))

```r
> library(psych)
> sessionInfo()
R Under development (unstable) (2020-02-21 r77847)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Catalina 10.15.3

Matrix products: default
BLAS: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib...

locale:

attached base packages:
[1] stats graphics grDevices utils datasets methods base

other attached packages:
[1] psychTools_2.0.1 psych_2.0.2

loaded via a namespace (and not attached):
[1] compiler_4.0.0 tools_4.0.0 parallel_4.0.0 foreign_0.8-75 nlme_3.1-144
mnormt_1.5-5 grid_4.0.0 lattice_0.20-40
```
R is extensible: The use of “packages”

- More than 15,514 packages are available for R (and growing daily)
- Can search all packages that do a particular operation by using the sos package
  - `install.packages("sos")` #if you haven’t already
  - `library(sos)` # make it active once you have it
    - `findFn("X")` #will search a web data base for all packages/functions that have "X"
    - `findFn("principal components analysis ")` #will return 2,554 matches and reports the top 400 and download 385 links to 174 packages
    - `findFn("Item Response Theory")` # will return 503 matches with 326 links in 76 packages
    - `findFn("INDSCAL ")` # will return 20 matches in 5 packages.
- `install.packages("X")` will install a particular package (add it to your R library – you need to do this just once)
- `library(X)` #will make the package X available to use if it has been installed (and thus in your library)
A small subset of very useful packages

- General use
  - core R
  - MASS
  - lattice (core)
  - lme4 (core)
  - psych
  - Zelig

- Special use
  - ltm
  - sem
  - lavaan
  - OpenMx
  - GPArotation
  - mnormt
  - > 15,514 known
  - + ?

- General applications
  - most descriptive and inferential stats
  - Modern Applied Statistics with S
  - Lattice or Trellis graphics
  - Linear mixed-effects models
  - Personality and psychometrics
  - General purpose toolkit

- More specialized packages
  - Latent Trait Model (IRT)
  - SEM and CFA (multiple groups)
  - SEM and CFA (multiple groups +)
  - Jennrich + Browne rotations
  - Multivariate distributions
  - Thousands of more packages on CRAN
  - Code on webpages/journal articles
Implementations

1. Base R in the Unix/Linux/Mac X11 framework
2. Base R on the Mac/PC
   - Mac has prompts at bottom of window
3. Graphical User Interfaces
   - R Commander
   - R studio as a convenient shell
**R Commander** (by John Fox) has a basic GUI.

---

R Commander is a GUI for R that provides an accessible interface for statistical analysis. It is designed to be user-friendly and suitable for users who are not familiar with the command-line interface of R. The software includes a comprehensive set of tools for data manipulation, statistical analysis, and graphical displays. The basic GUI of R Commander is equipped with a wide range of features, making it a valuable tool for both learning and application in statistical analysis.
RStudio (particularly nice for PCs)
Using R

1. Install the relevant packages (just once!)
   - Either one at a time, or by using a “task view”

2. Make the packages you want to use “active” by library(package name) e.g., library(psych), library(psychTools).
   - You need to do this for each session or it
   - Can be automatized at startup with
     .First() <- function() {library(psych); library(psychTools)}

3. Use the functions in a package
   - To see all functions in a package go to the index of the package or use the objects function: e.g., objects(package:psych)
   - Apply a function to data
   - All functions require an object to act upon. Most require this in parentheses. All functions return an object. This may be saved for later.
     - function(object) #apply the function to the object, show it
     - sqrt(2)
     - result <- function(object) #apply the function to object, save it
     - answer <- alpha(ability) #lots and lots of output is saved
What makes R useful?

1. The basic philosophy of open source allows one to see the code and modify it.
2. Object oriented meaning that all functions return objects that can be acted upon by other functions.
   - e.g., output of `mean` and `sd` is included in `describe` or `describeBy`
   - Output of `cor` is used in `fa`, `alpha`, etc.
   - Can add other functions (e.g., `tetrachoric` to do correlations instead of `cor`)
3. This is perhaps the most powerful feature of R!
4. R is vectorized, so that if it works for X, it will work for $\vec{X}$
Basic R commands – remember don’t enter the >

R is just a fancy calculator. Add, subtract, sum, products, group

```r
> 2 + 2
[1] 4

> 3^4
[1] 81

> sum(1:10)
[1] 55

> prod(c(1, 2, 3, 5, 7))
[1] 210
```

It is also a statistics table (the normal distribution, the t distribution, F, $\chi^2$, ...)

```r
> pnorm(q = 1) #probability of normal value > 1
[1] 0.8413447

> pt(q = 2, df = 20) #probability of t > 2 with 20 df
[1] 0.9703672
```
R is a set of distributions. Don’t buy a stats book with tables!

Table: To obtain the density, prefix with \( d \), probability with \( p \), quantiles with \( q \) and to generate random values with \( r \). (e.g., the normal distribution may be chosen by using `dnorm`, `pnorm`, `qnorm`, or `rnorm`.)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>base name</th>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>example application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>norm</td>
<td>mean</td>
<td>sigma</td>
<td></td>
<td>Most data</td>
</tr>
<tr>
<td>Multivariate normal</td>
<td>mvnorm</td>
<td>mean</td>
<td>( r )</td>
<td>sigma</td>
<td>Most data</td>
</tr>
<tr>
<td>Log Normal</td>
<td>lnorm</td>
<td>log mean</td>
<td>log sigma</td>
<td></td>
<td>income or reaction time</td>
</tr>
<tr>
<td>Uniform</td>
<td>unif</td>
<td>min</td>
<td>max</td>
<td></td>
<td>rectangular distributions</td>
</tr>
<tr>
<td>Binomial</td>
<td>binom</td>
<td>size</td>
<td>prob</td>
<td></td>
<td>Bernoulli trials (e.g., coin flips)</td>
</tr>
<tr>
<td>Student's t</td>
<td>t</td>
<td>df</td>
<td></td>
<td>nc</td>
<td>Finding significance of a t-test</td>
</tr>
<tr>
<td>Multivariate t</td>
<td>mvt</td>
<td>df</td>
<td>corr</td>
<td>nc</td>
<td>Multivariate applications</td>
</tr>
<tr>
<td>Fisher's ( F )</td>
<td>f</td>
<td>df1</td>
<td>df2</td>
<td>nc</td>
<td>Testing for significance of F test</td>
</tr>
<tr>
<td>( \chi^2 )</td>
<td>chisq</td>
<td>df</td>
<td></td>
<td>nc</td>
<td>Testing for significance of ( \chi^2 )</td>
</tr>
<tr>
<td>Exponential</td>
<td>exp</td>
<td>rate</td>
<td></td>
<td></td>
<td>Exponential decay</td>
</tr>
<tr>
<td>Gamma</td>
<td>gamma</td>
<td>shape</td>
<td>rate</td>
<td>scale</td>
<td>distribution theory</td>
</tr>
<tr>
<td>Hypergeometric</td>
<td>hyper</td>
<td>m</td>
<td>n</td>
<td>k</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>Logistic</td>
<td>logis</td>
<td>location</td>
<td>scale</td>
<td></td>
<td>Count data</td>
</tr>
<tr>
<td>Poisson</td>
<td>pois</td>
<td>lambda</td>
<td></td>
<td></td>
<td>Reaction time distributions</td>
</tr>
<tr>
<td>Weibull</td>
<td>weibull</td>
<td>shape</td>
<td>scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cauchy</td>
<td>cauchy</td>
<td>location</td>
<td>scale</td>
<td>log</td>
<td>infinite variance!</td>
</tr>
</tbody>
</table>
R can draw distributions

curve(dnorm(x), -3, 3, ylab = "probability of x", main = "A normal curve")
R can draw more interesting distributions

Log normal

Chi Square distribution

Normal and t with 4 df

The normal curve
R is also a graphics calculator

The first line draws the normal curve, the second prints the title, the next lines draw the cross hatching.

```r
op <- par(mfrow=c(2,2))  # set up a 2 x 2 graph
curve(dnorm(x),-3,3,xlab="",ylab="Probability of z")
title(main="The normal curve",outer=FALSE)
xvals <- seq(-3,-2,length=100)
dvals <- dnorm(xvals)
polygon(c(xvals,rev(xvals)),c(rep(0,100),rev(dvals)),density=2,angle=-45)
xvals <- seq(-2,-1,length=100)
dvals <- dnorm(xvals)
polygon(c(xvals,rev(xvals)),c(rep(0,100),rev(dvals)),density=14,angle=45)
xvals <- seq(-1,0,length=100)
dvals <- dnorm(xvals)
polygon(c(xvals,rev(xvals)),c(rep(0,100),rev(dvals)),density=34,angle=-45)
xvals <- seq(2,3,length=100)
dvals <- dnorm(xvals)
polygon(c(xvals,rev(xvals)),c(rep(0,100),rev(dvals)),density=2,angle=45)
xvals <- seq(1,2,length=100)
dvals <- dnorm(xvals)
polygon(c(xvals,rev(xvals)),c(rep(0,100),rev(dvals)),density=14,angle=-45)
xvals <- seq(0,1,length=100)
dvals <- dnorm(xvals)
polygon(c(xvals,rev(xvals)),c(rep(0,100),rev(dvals)),density=34,angle=45)
curve(dlnorm(x),0,5,ylab='Probiabiity of log(x)',main='Log normal')
curve(dchisq(x,1),0,5,ylab='Probability of Chi Sq',xlab='Chi Sq',main='Chi Square distribution')
curve(dnorm(x),-4,4,ylab='Probability of z or t',xlab='z or t',main='Normal and t with 4 df')
curve(dt(x,4),add=TRUE)
op <- par(mfrow=c(1,1))
```
## Example data sets built into many packages

**Table:** Some of the 50 data sets in the psych or psychTools packages

<table>
<thead>
<tr>
<th>Name</th>
<th>Content description</th>
</tr>
</thead>
<tbody>
<tr>
<td>veg</td>
<td>Thurstone’s Vegetables</td>
</tr>
<tr>
<td>cities</td>
<td>Airplane distances for 11 US cities</td>
</tr>
<tr>
<td>galton</td>
<td>Francis Galton’s original data set of heights</td>
</tr>
<tr>
<td>cushny</td>
<td>The original t-test data from “student” (Gossett)</td>
</tr>
<tr>
<td>ability</td>
<td>16 ability items from SAPA</td>
</tr>
<tr>
<td>bfi</td>
<td>25 Big Five items + gender, age, education from SAPA</td>
</tr>
<tr>
<td>sat.act</td>
<td>Test scores, gender, age and education</td>
</tr>
<tr>
<td>Thurstone</td>
<td>9 ability variables from Thurstone</td>
</tr>
<tr>
<td>msq</td>
<td>75 mood items from the PMC lab</td>
</tr>
<tr>
<td>neo</td>
<td>Correlation matrix of the 30 NEO-PI-R facets</td>
</tr>
</tbody>
</table>

data() #to see all available
data(package=c("psych","psychTools")) # to see all psych data sets
A simple scatter plot using `plot` shows Fisher’s Iris data set

```
plot(iris[1:2], xlab="Sepal.Length", ylab="Sepal.Width", main="Fisher Iris data")
```
A scatter plot matrix with loess regression using `pairs.panels` shows more information than a simple scatter plot

1. Correlations above the diagonal
2. Diagonal shows histograms and densities
3. Scatter plots below the diagonal with correlation ellipse
4. Locally smoothed (loess) regressions for each pair

```r
pairs.panels(iris[1:4])
```
A better SPLOM with colors for groups using `pairs.panels`

1. Correlations above the diagonal
2. Diagonal shows histograms and densities
3. Scatter plots below the diagonal with correlation ellipse
4. Locally smoothed (loess) regressions for each pair
5. Optional color coding of grouping variables.

`pairs.panels(iris[1:4], bg=c("red","yellow","blue"), iris$Species, pch=21, main="Fisher Iris data by Species")`
Using R for psychological statistics: Basic statistics

1. Writing syntax
   • For a single line, just type it
   • Mistakes can be redone by using the up arrow key
   • For longer code, use a text editor (built into some GUIs)

2. Data entry
   • Using built in data sets for examples
   • Copying from another program (using the "clipboard")
   • Reading a text or csv file
   • Importing from SPSS or SAS
   • Simulate it (using various simulation routines)

3. Descriptives
   • Graphical displays
   • Descriptive statistics
   • Correlation

4. Inferential
   • the t test
   • the F test
   • the linear model
Data entry overview

1. Using built in data sets for examples
   - `data()` will list > 100 data sets in the datasets package as well as all sets in loaded packages.
   - Most packages have associated data sets used as examples
   - `psych` and `psychTools` have > 50 example data sets

2. Copying from another program
   - use copy and paste into R using `read.clipboard` and its variations

3. Reading a text or csv file
   - read a local or remote file using the `read.file` function

4. Importing from SPSS or SAS

5. Simulate it (using various simulation routines)
Examples of built in data sets from the psych package

```r
> data(package="psych")
Bechtoldt
Seven data sets showing a bifactor solution.
Dwyer
8 cognitive variables used by Dwyer for an example.
Reise
Seven data sets showing a bifactor solution.
all.income (income)
US family income from US census 2008
bfi
25 Personality items representing 5 factors
blot
Bond's Logical Operations Test - BLOT
burt
11 emotional variables from Burt (1915)
cities
Distances between 11 US cities
epi.bfi
13 personality scales from the Eysenck Personality Inventory and Big 5 inventory
affect
Two data sets of affect and arousal scores as a function of personality and movie conditions.
galton
Galton's Mid parent child height data
income
US family income from US census 2008
iqitems
16 multiple choice IQ items
msq
75 mood items from the Motivational State Questionnaire for N = 3896
neo
NEO correlation matrix from the NEO_PI_R manual
sat.act
3 Measures of ability: SATV, SATQ, ACT
Thurstone
The classic Thurstone 9 variable problem
veg (vegetables)
Paired comparison of preferences for 9 vegetables
```
Reading data from another program – using the clipboard

1. Read the data in your favorite spreadsheet or text editor
2. Copy to the clipboard
3. Execute the appropriate `read.clipboard` function with or without various options specified

   ```r
   my.data <- read.clipboard()  # assumes headers and tab or space delimited
   my.data <- read.clipboard.csv()  # assumes headers and comma delimited
   my.data <- read.clipboard.tab()  # assumes headers and tab delimited
   # (e.g., from Excel)
   my.data <- read.clipboard.lower()  # read in a matrix given the lower
   my.data <- read.clipboard.upper()  # or upper off diagonal
   my.data <- read.clipboard.fwf()  # read in data using a fixed format width
   # (see `read.fwf` for instructions)
   ```

4. `read.clipboard()` has default values for the most common cases and these do not need to be specified. Consult `?read.clipboard` for details. (Note, the `read.clipboard` and `read.file` functions require the `psychTools`. Install and make it active. e.g., `library(psychTools)`.)
Reading from a local or remote file

1. Perhaps the standard way of reading in data is using the `read` command.
   - First must specify the location of the file
   - Can either type this in directly or use the `file.choose` function
   - The file name/location can be a remote URL

2. Two examples of reading data

   ```r
   file.name <- file.choose()  # this opens a window to allow you find the file
   my.data <- read.table(file.name)  # read the file
   my.data <- read.file()  # opens a search window, reads the file according to the suffix
   datafilename="http://personality-project.org/r/datasets/R.appendix1.data"
   data.ex1=read.table(datafilename,header=TRUE)  # read the data into a table
   > dim(data.ex1)  # what are the dimensions of what we read?
   [1] 18 2
   > describe(data.ex1)  # do the data look right?
   var  n mean  sd median trimmed mad min max range skew kurtosi
    Dosage*   1 18 1.89 0.76   2  1.88 1.48  1  3  2 0.16 -1.12
   Alertness 2 18 27.67 6.82  27 27.50 8.15 17 41 24 0.25 -0.68
   ```
read a “foreign” file e.g., an SPSS sav file

read.spss reads a file stored by the SPSS save or export commands. Can be done using the read.file command

read.spss(file, use.value.labels = TRUE, to.data.frame = FALSE, max.value.labels = Inf, trim.factor.names = FALSE, trim_values = TRUE, reencode = NA, use.missings = to.data.frame)

- **file**: Character string: the name of the file or URL to read.
- **use.value.labels**: Convert variables with value labels into R factors with those levels?
- **to.data.frame**: return a data frame? Defaults to FALSE, probably should be TRUE in most cases.
- **max.value.labels**: Only variables with value labels and at most this many unique values will be converted to factors if use.value.labels = TRUE.
- **trim.factor.names**: Logical: trim trailing spaces from factor levels?
- **trim_values**: logical: should values and value labels have trailing spaces ignored when matching for use.value.labels = TRUE?
- **use.missings**: logical: should information on user-defined missing values be used to set the corresponding values to NA?
Simulate data

For many demonstration purposes, it is convenient to generate simulated data with a certain defined structure. The *psych* package has a number of built in simulation functions. Here are a few of them.

1. Simulate various item structures
   - `sim.congeneric` A one factor congeneric measure model
   - `sim.items` A two factor structure with either simple structure or a circumplex structure.
   - `sim.rasch` Generate items for a one parameter IRT model.
   - `sim.irt` Generate items for a one-four parameter IRT Model
2. Simulate various factor structures
   - `sim.simplex` Default is a four factor structure with a three time point simplex structure.
   - `sim.hierarchical` Default is 9 variables with three correlated factors.
Get the data and look at it

Read in some data, look at the first and last few cases, and then get basic descriptive statistics. For this example, we will use a built in data set.

```r
> my.data <- epi.bfi
> headtail(my.data)
```

```
epiE epiS epiImp epilie epiNeur bfagree bfcon bfext bfneur bfopen bdi traitanx stateanx
1   18   10    7    3    9   138    96   141    51  138    1    24    22
2   16    8    5    1   12   101    99  107   116  132    7    41    40
3    6    1    3    2    5  143   118   38   68   90    4    37    44
4   12    6    4    3   15  104   106   64  114  101    8    54    40
```

... ... ... ... ... ... ... ... ... ... ... ... ...

```
228  12    7    4    3   15  155   129  127    88  110    9    35    34
229  19   10    7    2   11  162  152  163  104  164    1    29    47
230    4    1    1    2   10   95  111   75  123  138    5    39    58
231    8    6    3    2   15   85   62   90  131    96    24    58    58
```

epi.bfi has 231 cases from two personality measures
Now find the descriptive statistics for this data set

```r
> describe(my.data)

<table>
<thead>
<tr>
<th>var</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>median</th>
<th>trimmed</th>
<th>mad</th>
<th>min</th>
<th>max</th>
<th>range</th>
<th>skew</th>
<th>kurtosis</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>epiE</td>
<td>1</td>
<td>13.33</td>
<td>4.14</td>
<td>14</td>
<td>13.49</td>
<td>4.45</td>
<td>1</td>
<td>22</td>
<td>21</td>
<td>-0.33</td>
<td>-0.01</td>
<td>0.27</td>
</tr>
<tr>
<td>epiS</td>
<td>2</td>
<td>7.58</td>
<td>2.69</td>
<td>8</td>
<td>7.77</td>
<td>2.97</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td>-0.57</td>
<td>0.04</td>
<td>0.18</td>
</tr>
<tr>
<td>epiImp</td>
<td>3</td>
<td>4.37</td>
<td>1.88</td>
<td>4</td>
<td>4.36</td>
<td>1.48</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>0.06</td>
<td>-0.59</td>
<td>0.12</td>
</tr>
<tr>
<td>epilie</td>
<td>4</td>
<td>2.38</td>
<td>1.50</td>
<td>2</td>
<td>2.27</td>
<td>1.48</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0.66</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>epiNeur</td>
<td>5</td>
<td>10.41</td>
<td>4.90</td>
<td>10</td>
<td>10.39</td>
<td>4.45</td>
<td>0</td>
<td>23</td>
<td>23</td>
<td>0.06</td>
<td>-0.59</td>
<td>0.12</td>
</tr>
<tr>
<td>bfagree</td>
<td>6</td>
<td>125.00</td>
<td>18.14</td>
<td>126</td>
<td>125.26</td>
<td>17.79</td>
<td>74</td>
<td>167</td>
<td>93</td>
<td>-0.21</td>
<td>-0.22</td>
<td>1.19</td>
</tr>
<tr>
<td>bfcon</td>
<td>7</td>
<td>113.25</td>
<td>21.88</td>
<td>114</td>
<td>113.42</td>
<td>22.24</td>
<td>53</td>
<td>178</td>
<td>125</td>
<td>-0.02</td>
<td>0.29</td>
<td>1.44</td>
</tr>
<tr>
<td>bfext</td>
<td>8</td>
<td>102.18</td>
<td>26.45</td>
<td>104</td>
<td>102.99</td>
<td>22.24</td>
<td>8</td>
<td>168</td>
<td>160</td>
<td>-0.41</td>
<td>0.58</td>
<td>1.74</td>
</tr>
<tr>
<td>bfneur</td>
<td>9</td>
<td>87.97</td>
<td>23.34</td>
<td>90</td>
<td>87.70</td>
<td>23.72</td>
<td>34</td>
<td>152</td>
<td>118</td>
<td>0.07</td>
<td>-0.51</td>
<td>1.54</td>
</tr>
<tr>
<td>bfopen</td>
<td>10</td>
<td>123.43</td>
<td>20.51</td>
<td>125</td>
<td>123.78</td>
<td>20.76</td>
<td>73</td>
<td>173</td>
<td>100</td>
<td>-0.16</td>
<td>-0.11</td>
<td>1.35</td>
</tr>
<tr>
<td>bdi</td>
<td>11</td>
<td>6.78</td>
<td>5.78</td>
<td>6</td>
<td>5.97</td>
<td>4.45</td>
<td>0</td>
<td>27</td>
<td>27</td>
<td>1.29</td>
<td>1.60</td>
<td>0.38</td>
</tr>
<tr>
<td>traitanx</td>
<td>12</td>
<td>39.01</td>
<td>9.52</td>
<td>38</td>
<td>38.36</td>
<td>8.90</td>
<td>22</td>
<td>71</td>
<td>49</td>
<td>0.67</td>
<td>0.54</td>
<td>0.63</td>
</tr>
<tr>
<td>stateanx</td>
<td>13</td>
<td>39.85</td>
<td>11.48</td>
<td>38</td>
<td>38.92</td>
<td>10.38</td>
<td>21</td>
<td>79</td>
<td>58</td>
<td>0.72</td>
<td>0.04</td>
<td>0.76</td>
</tr>
</tbody>
</table>
```

describe is one of the most important functions to use on any new data set. Look at the ranges, the minima, the maxima, the sds.
Boxplots are a convenient descriptive device

Show the Tukey “boxplot” for the Eysenck Personality Inventory

`boxplot(my.data[1:5])` #just the first 5 variables

Boxplots of EPI scales
Enhanced box plots are even more convenient descriptive devices

Show the Tukey “boxplot” for the Eysenck Personality Inventory

```r
boxplot(my.data[1:5]) #just the first 5 variables
```

A notched boxplot of the epi
A violin plot shows the density distribution more graphically

Density plot

Observed

epiE epiS epiImp epilie epiNeur

0 5 10 15 20
Plot the scatter plot matrix (SPLOM) of the first 5 variables using the `pairs.panels` function.

Use the `pairs.panels` function from `psych`.

```r
pairs.panels(my.data[1:5])
```

Note that we specify the range of variables.
Find the correlations for this data set, round off to 2 decimal places

```r
> round(cor(my.data, use = "pairwise"), 2)
```

<table>
<thead>
<tr>
<th></th>
<th>epiE</th>
<th>epiS</th>
<th>epiImp</th>
<th>epilie</th>
<th>epiNeur</th>
<th>bfagree</th>
<th>bfcon</th>
<th>bfext</th>
<th>bfneur</th>
<th>bfopen</th>
<th>bdi</th>
<th>traitanx</th>
<th>stateanx</th>
</tr>
</thead>
<tbody>
<tr>
<td>epiE</td>
<td>1.00</td>
<td>0.85</td>
<td>0.80</td>
<td>-0.22</td>
<td>-0.18</td>
<td>0.18</td>
<td>-0.11</td>
<td>0.54</td>
<td>-0.09</td>
<td>0.14</td>
<td>-0.16</td>
<td>-0.23</td>
<td>-0.13</td>
</tr>
<tr>
<td>epiS</td>
<td>0.85</td>
<td>1.00</td>
<td>0.43</td>
<td>-0.05</td>
<td>-0.22</td>
<td>0.20</td>
<td>0.05</td>
<td>0.58</td>
<td>-0.07</td>
<td>0.15</td>
<td>-0.13</td>
<td>-0.26</td>
<td>-0.12</td>
</tr>
<tr>
<td>epiImp</td>
<td>0.80</td>
<td>0.43</td>
<td>1.00</td>
<td>-0.24</td>
<td>-0.07</td>
<td>0.08</td>
<td>-0.24</td>
<td>0.35</td>
<td>-0.09</td>
<td>0.07</td>
<td>-0.11</td>
<td>-0.12</td>
<td>-0.09</td>
</tr>
<tr>
<td>epilie</td>
<td>-0.22</td>
<td>-0.05</td>
<td>0.24</td>
<td>1.00</td>
<td>-0.25</td>
<td>0.17</td>
<td>0.23</td>
<td>-0.04</td>
<td>-0.22</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.23</td>
<td>-0.15</td>
</tr>
<tr>
<td>epiNeur</td>
<td>-0.18</td>
<td>-0.22</td>
<td>0.07</td>
<td>-0.25</td>
<td>1.00</td>
<td>-0.08</td>
<td>-0.13</td>
<td>-0.17</td>
<td>0.63</td>
<td>0.09</td>
<td>0.63</td>
<td>0.67</td>
<td>0.57</td>
</tr>
<tr>
<td>bfagree</td>
<td>0.18</td>
<td>0.20</td>
<td>0.08</td>
<td>0.17</td>
<td>-0.08</td>
<td>1.00</td>
<td>0.45</td>
<td>0.48</td>
<td>-0.04</td>
<td>0.39</td>
<td>0.47</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>bfcon</td>
<td>-0.11</td>
<td>0.05</td>
<td>-0.24</td>
<td>0.23</td>
<td>-0.13</td>
<td>0.45</td>
<td>1.00</td>
<td>0.27</td>
<td>0.04</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>bfext</td>
<td>0.54</td>
<td>0.58</td>
<td>0.35</td>
<td>-0.04</td>
<td>-0.17</td>
<td>0.48</td>
<td>0.27</td>
<td>1.00</td>
<td>0.04</td>
<td>0.46</td>
<td>0.46</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>bfneur</td>
<td>-0.09</td>
<td>-0.07</td>
<td>0.09</td>
<td>-0.22</td>
<td>0.63</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>1.00</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>bfopen</td>
<td>0.14</td>
<td>0.15</td>
<td>0.07</td>
<td>-0.03</td>
<td>0.09</td>
<td>0.39</td>
<td>0.31</td>
<td>0.46</td>
<td>0.29</td>
<td>1.00</td>
<td>0.00</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>bdi</td>
<td>-0.16</td>
<td>-0.13</td>
<td>-0.11</td>
<td>0.20</td>
<td>0.58</td>
<td>-0.14</td>
<td>-0.18</td>
<td>-0.14</td>
<td>0.47</td>
<td>-0.08</td>
<td>1.00</td>
<td>0.09</td>
<td>0.19</td>
</tr>
<tr>
<td>traitanx</td>
<td>-0.23</td>
<td>-0.26</td>
<td>-0.12</td>
<td>-0.23</td>
<td>0.73</td>
<td>-0.31</td>
<td>-0.29</td>
<td>-0.39</td>
<td>0.59</td>
<td>-0.11</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>stateanx</td>
<td>-0.13</td>
<td>-0.12</td>
<td>-0.09</td>
<td>-0.15</td>
<td>0.49</td>
<td>-0.19</td>
<td>-0.14</td>
<td>-0.15</td>
<td>0.49</td>
<td>-0.04</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Find the correlations using `lowerCor`

```r
> lowerCor(my.data)
```

<table>
<thead>
<tr>
<th></th>
<th>epiE</th>
<th>epiS</th>
<th>epImp</th>
<th>epili</th>
<th>epiNr</th>
<th>bfagr</th>
<th>bfcon</th>
<th>bfext</th>
<th>bfner</th>
<th>bfopen</th>
<th>bdi</th>
<th>traitanx</th>
<th>stateanx</th>
</tr>
</thead>
<tbody>
<tr>
<td>epiE</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epiS</td>
<td>0.85</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epImp</td>
<td>0.80</td>
<td>0.43</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epili</td>
<td>-0.22</td>
<td>-0.05</td>
<td>-0.24</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epiNr</td>
<td>-0.18</td>
<td>-0.22</td>
<td>-0.07</td>
<td>-0.25</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bfagr</td>
<td>0.18</td>
<td>0.20</td>
<td>0.08</td>
<td>0.17</td>
<td>-0.08</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bfcon</td>
<td>-0.11</td>
<td>0.05</td>
<td>-0.24</td>
<td>0.23</td>
<td>-0.13</td>
<td>0.45</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bfext</td>
<td>0.54</td>
<td>0.58</td>
<td>0.35</td>
<td>-0.04</td>
<td>-0.17</td>
<td>0.48</td>
<td>0.27</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bfner</td>
<td>-0.09</td>
<td>-0.07</td>
<td>-0.09</td>
<td>-0.22</td>
<td>0.63</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bfopen</td>
<td>0.14</td>
<td>0.15</td>
<td>0.07</td>
<td>-0.03</td>
<td>0.09</td>
<td>0.39</td>
<td>0.31</td>
<td>0.46</td>
<td>0.29</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bdi</td>
<td>-0.16</td>
<td>-0.13</td>
<td>-0.11</td>
<td>-0.20</td>
<td>0.58</td>
<td>-0.14</td>
<td>-0.18</td>
<td>-0.14</td>
<td>0.47</td>
<td>-0.08</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>traitanx</td>
<td>-0.23</td>
<td>-0.26</td>
<td>-0.12</td>
<td>-0.23</td>
<td>0.73</td>
<td>-0.31</td>
<td>-0.29</td>
<td>-0.39</td>
<td>0.59</td>
<td>-0.11</td>
<td>0.65</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>stateanx</td>
<td>-0.13</td>
<td>-0.12</td>
<td>-0.09</td>
<td>-0.15</td>
<td>0.49</td>
<td>-0.19</td>
<td>-0.14</td>
<td>-0.15</td>
<td>0.49</td>
<td>-0.04</td>
<td>0.61</td>
<td>0.59</td>
<td></td>
</tr>
</tbody>
</table>
A heat map of 25 BFI items using cor.plot

Big 5 Inventory Items from SAPA


