Why Should We Care About Measurement?: Reliability Validity, and Scaling

William F. Chaplin, Ph.D.
Department of Psychology
St John’s University
New York

Measurement is the Foundation of Science

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 meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.

----- William Thomson, Lord Kelvin, Popular Lectures and Addresses [1891-1894]

Measure what is measurable, and make measurable what is not so.

-----Gottlob Frege (1848 - 1925)

Nothing has really happened until it has been recorded.

------Virginia Woolf

Through and through the world is infested with quantity:
To talk sense is to talk quantities. It is no use saying the nation is large .. How large? It is no use saying the radium is scarce … How scarce? You cannot evade quantity. You may fly to poetry and music, and quantity and number will face you in your rhythms and your octaves.

-------- Alfred North Whitehead

But Measurement is Difficult and Inexact

All exact science is dominated by the idea of approximation.

------ Bertrand Russell

There must be an ideal world, a sort of mathematician's paradise where everything happens as it does in textbooks.

--------Bertrand Russell

“Intelligence is whatever intelligence tests measure”

-------Boring, E.G. (1923) Intelligence as the tests test it. *The New Republic*, June, 35-189.

And Measurement is Unexciting

"Measurement is the Achilles' heel of sociobehavioral research.
Although most programs in sociobehavioral sciences ... require a medium
of exposure to statistics and research design, few seem
to require the same where measurement is concerned ...
It is, therefore, not surprising that little or no attention is given
to the properties of the measures used in many research studies."

-------Pedhazur, E.J., & Schmelkin, L.P. (1991) Measurement, Design
and Analysis: An Integrated Approach. Hillsdale NJ: Erlbaum. (p. 2-3).

“Yet, when we teach courses on measurement and test construction, we seldom encounter much enthusiasm. In fact, most students think that measurement is outright boring.”

------John & Benet-Martinez (2000)

Why Is Measurement Important for Our Research?

>Type II error

(power)

>Type III error

(doing the wrong study)

>Interpreting model parameters

(clinical and substantive as opposed to statistical significance)

Reliability and Type II error

Measurement unreliability attenuates the obtained relations among variables.

$$\rho_{xy}\sqrt{r_{xx}}\sqrt{r_{yy}} = r_{xy}$$

True relation

r
e

Reliability of x

Reliability of y

Observed relation

Example:

The true correlation between treatment and outcome is .20

(e.g. Clozapine vs conventional neuroleptics and clinical improvement in schizophrenia (Wahlbeck et al., 1999)

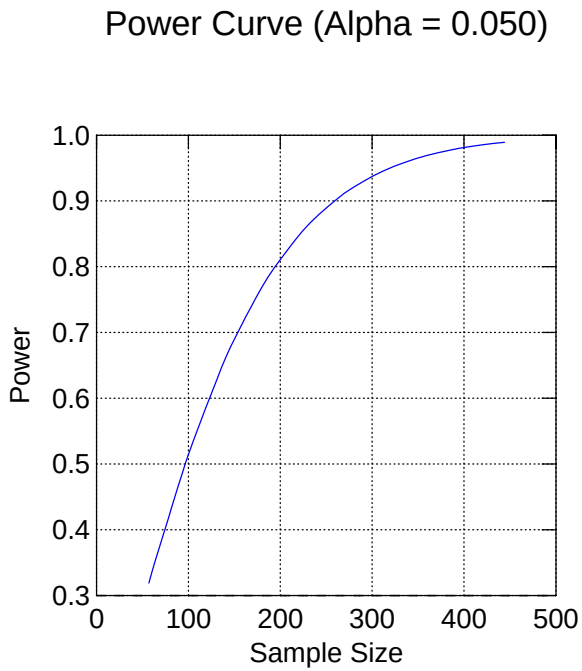
The reliability of treatment group assignment is 1.0

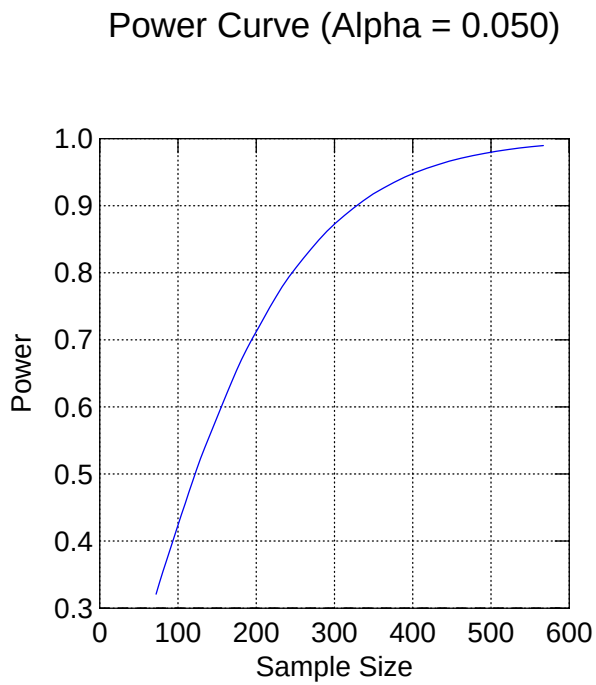
If the reliability of the measure of clinical improvement is .80 then we will observe a treatment – outcome relation of .178; if the reliability is .70 then the observed relation will be .167 and so on.

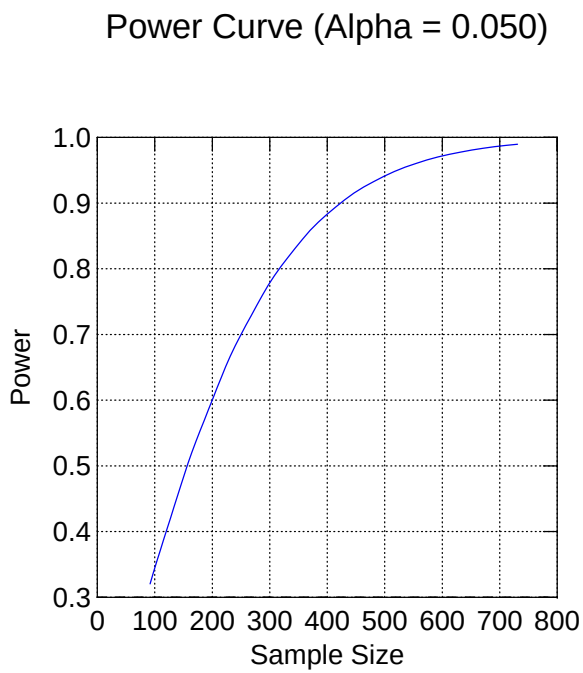
Note that if treatment group assignment contains some error (say it is 90% accurate) then the observed correlations will be .168 and .157 respectively.

Implications for Statistical Power

Power set at .80







r = .20	r = .178	r = .157
Needed N = 194	246	316

(actual study N = 1,850)

Correction for Attenuation

$$\rho_{xy} = r_{xy} / \left(\sqrt{rxx} \sqrt{ryy} \right)$$

What do we do in our study design that hurts reliability? (and power)

Use single item measures or shorten existing measures to save subject time or perhaps money.

Shortening measures has a calculable effect on reliability.

Spearman-Brown “prophecy” formula

$$r'_{xx} = k(r_{xx}) / [1 + (k-1)(r_{xx})]$$

Where:

r'_{xx} is the estimated new reliability

r_{xx} is the obtained reliability of the original measure

k is the ratio of the number of items on the new measure relative to the old measure (I_{new} / I_{old})

Example

The reliability of the 40 item Narcissism Personality Inventory (NPI) is reported to be .80.

An investigator studying the relation between the NPI and Aggression decides to use only a five item version of the NPI.

The reliability of this five item version is estimated to be

$$[(5/40)(.80)]/[1 + (5/40-1)(.80)] = .33$$

(the reliability of the aggression measure is reported to be .83)

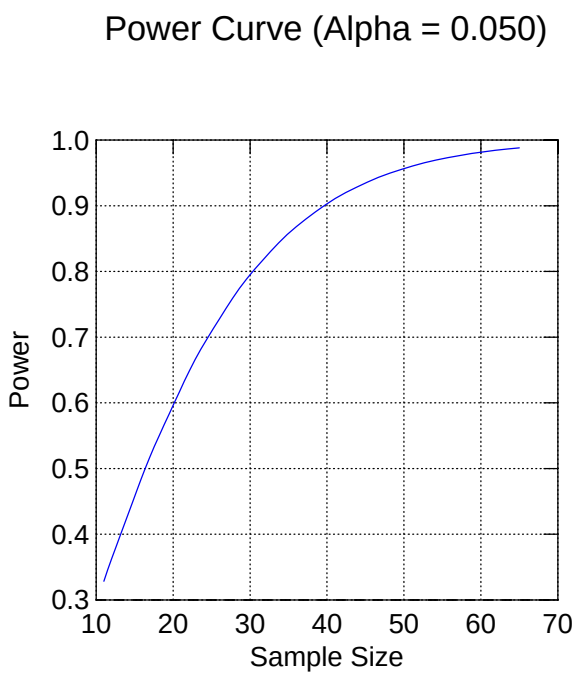
Attenuating Effect on the Narcissism-Aggression Relation

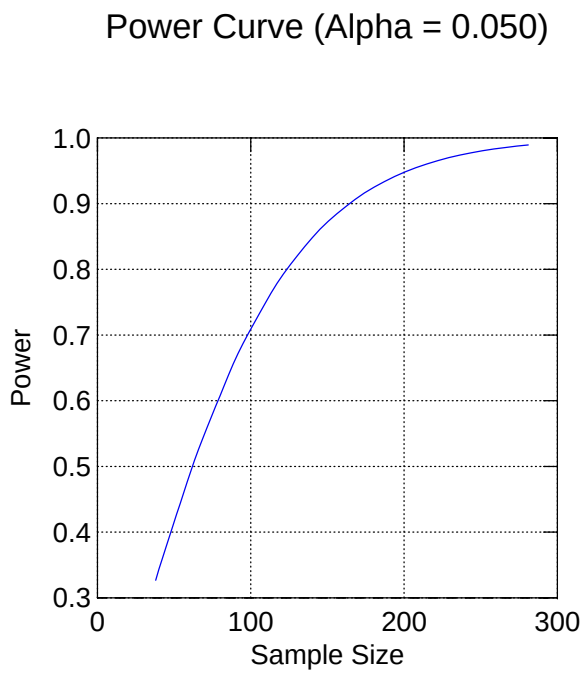
The hypothesized relation between aggression and narcissism (based on perfect measures) is .49.

This can be expected to be attenuated using the unreliable measures.

$$.49(.91)(.57) = .25$$

Effect on Power Analysis for power of .80





r = .49

N = 31

r = .25

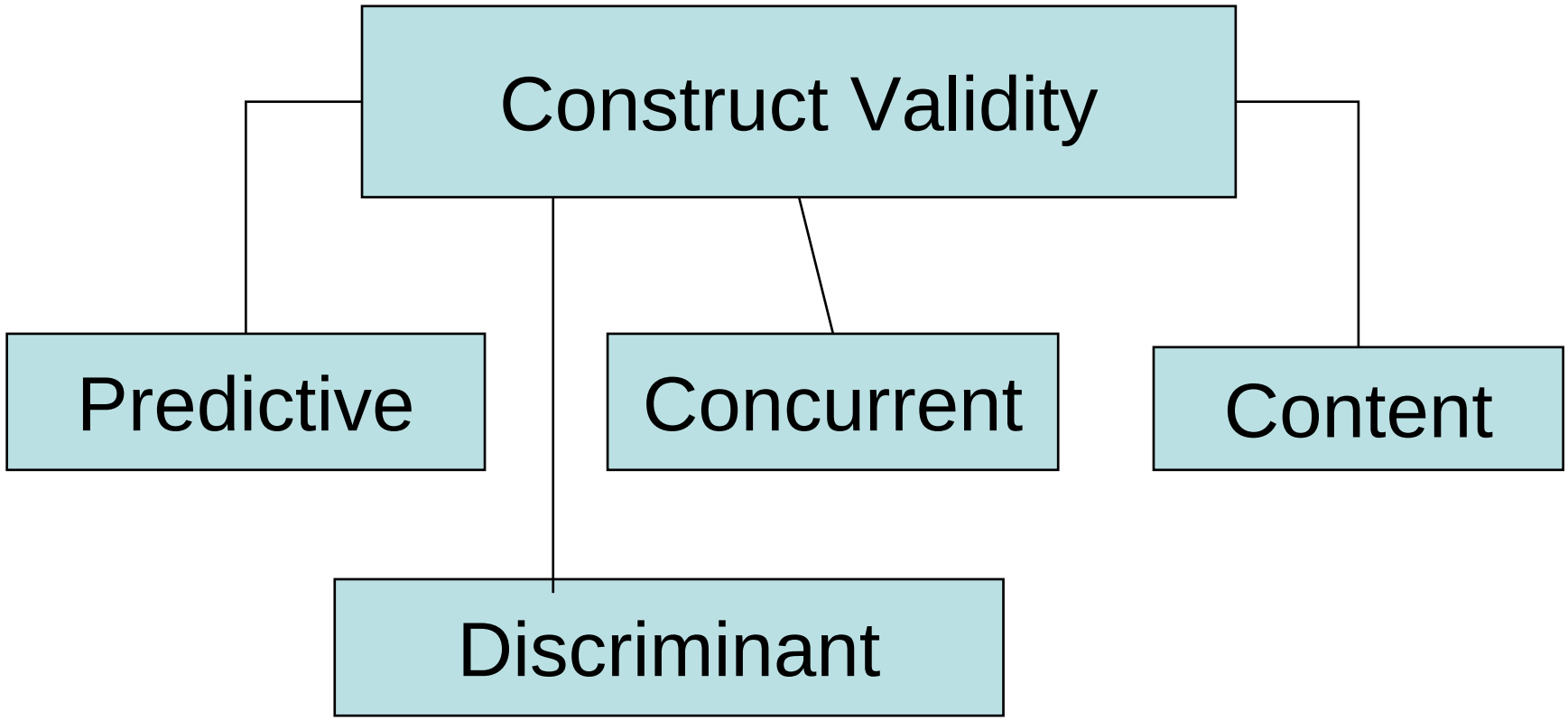
N = 124

Validity and Type III error

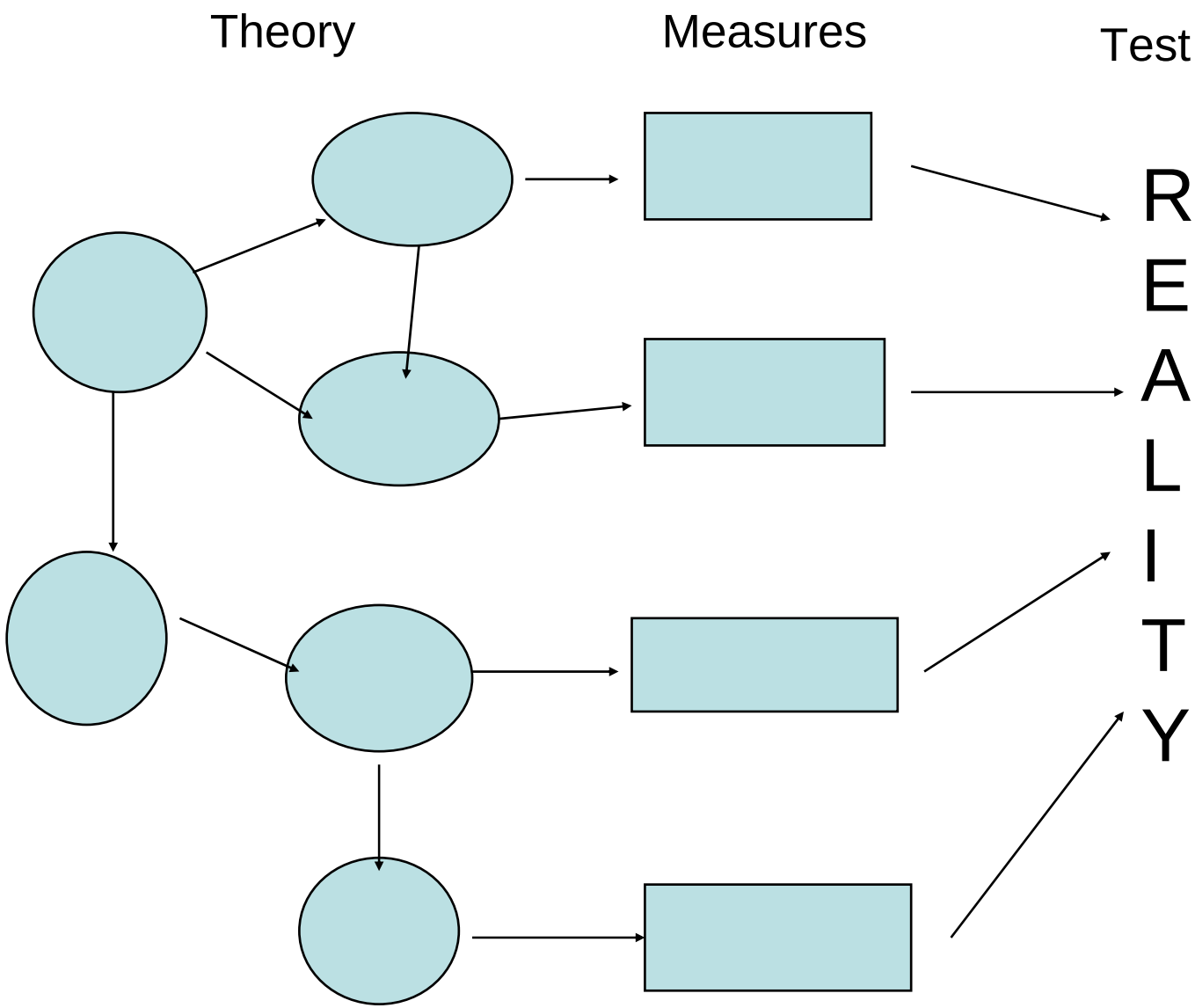
Type III error has been used to refer to a variety of mistakes, but I am using it here to refer to “solving the wrong problem correctly” or “correctly answering the wrong question.”

In this case the wrong question is caused by using an invalid measure: A measure that represents something other than what we think it does.

Validity


Construct Validity
Predictive
Concurrent
Content
Discriminant

Nomological Network Cronbach and Meehl (1955)



Circularity of Substantive and Measurement Theories

1) X is a measure of A

2) Y is a measure of B

3) A and B are negatively correlated

4) Are X & Y negatively correlated?

Hopefully the circularity is an upward spiral!

Content Validity

Item Content Provides an Ostensive Definition of a Construct

There is no good quantitative index of content validity

But content can provide clues to interpreting study results. Did we measure what we thought we did?

Loneliness as an Example

"Because most of the self report measures [of loneliness] for children contain diverse item content that goes beyond loneliness per se (as does the widely used UCLA Loneliness Scale for adults), caution must be used when interpreting the results. Some investigators have therefore calculated 'pure loneliness' scores by using only items that directly assess feelings of loneliness."

Asher, S. R. & Paquette, J. A, (2003) Loneliness and peer relations in children. *Current Directions in Psychological Science, 12,* 75-78.

UCLA Loneliness Scale (Version 3)

Russell, D. W. (1996). Loneliness scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment, 66,* 20-40.

Response is on a 4 point scale 1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Always.

How often do you feel…..

1) that you are in tune with the people around you?

2) that you lack companionship?

3) that there is no one you can turn to?

4) alone?

5. part of a group of friends?
6. that you have a lot in common with the people around you?
7. that you are no longer close to anyone?
8. that your interests and ideas are not shared by those around you?
9. outgoing and friendly?
10. close to people?
11. left out?
12. that your relationships with others are not meaningful?
13. that no one really knows you well?
14. isolated from others?

15. that you can find companionship when you want it?
16. that there are people who really understand you?
17. shy?
18. that there are people around you but not with you?
19. that there are people you can talk to?
20. that there are people you can turn to?

NYU Loneliness Scale

Rubenstein and Shaver (1982)

I am a lonely person.

How often do you feel lonely

I will always be a lonely person

Other people think of me as lonely

I always was a lonely person

Compared to other people how lonely do you think you are?

When I am completely alone, I feel lonely?

When you feel lonely how lonely do you feel?

Loneliness Scale

de jong, G. J., & Kamphuis, F. H. (1985). The development of a Rasch-type loneliness scale. *Applied Psychological Measurement, 9,* 289-299.

1. There is always someone I can talk to about my day to day problems.
2. I miss having a really close friend.
3. I experience a general sense of emptiness.
4. There are plenty of people I can lean on when I have problems.
5. I miss the pleasure of the company of others

6. I find my circle of friends too limited.

7. There are many people I can trust completely.

8. There are enough people I feel close to.

9. I miss having people around.

10. I often feel rejected.

11. I can call on my friends whenever I need them.

Differential Loneliness Scale

Schmidt, N., & Sermat, V. (1983). Measuring loneliness in different relationships. *Journal of Personality and Social Psychology, 44,* 1038-1047.

Family

I feel close to members of my family.

I have little contact with members of my family

I do not get along well with members of my family

I have a good relationship with most members of my immediate family

My family seldom really listens to what I say.

Romantic/Sexual

I have a lover or spouse with whom I can discuss my important problems and worries

I am now involved in a romantic or marital relationship where both of us are making a genuine effort at cooperation

My lover or spouse sense when I am troubled or encourages me

I feel valued and respected in my current romantic or marital relationship.

I seldom get the emotional security I need from a good romantic or sexual relationship

Friends

I do not feel that I can turn to my friends living around me for help when I need it.

I allow myself to become close to my friends

I do not have many friends in the city where I live

I get plenty of help and support from my friends

Few of my friends understand me the way I want to be understood.

Groups/Community

I feel I really do not have much in common with the larger community in which I live.

No one in the community where I live seems to care much about me

I feel that I have “roots” (sense of belonging) in the larger community or neighborhood I live in.

I do not have any neighbors who would help me out in a time of need

I know people in my community who understand and share my views and beliefs

Summary

I am not convinced that any of these measures provide a pure index of loneliness.

Social support

Shyness

Introversion

Neuroticism

Solutions

1. Spend more time developing measures.

Measure development IS Theory Testing

2. Revise the measure by deleting items
3. Partial offending constructs from the measure

Scaling and the Interpretation of Model Parameters

It is no longer enough to claim that an effect (model parameter) is “significant.” We also need to interpret that effect substantively.

Effect size, clinical significance, odds ratios, dose response etc.

The values of model parameters depends not only on their significance, but on the variables are scaled (measured).

> In statistical analysis, the meaning (i.e. measurement or scaling) of the variables is often irrelevant.

> In substantive analysis the scaling of the variables is crucial.

> Why?

> In statistical analysis we focus on the statistical significance of parameters and factors that effect the legitimacy of those statistical tests.

> In substantive analysis the meaning of the parameters is crucial. And the scaling of variables has implications for the meaning of the parameters.

An Example: Qualitative (categorical) Variables

Four experimental groups

1 = control

2 = low exposure

3 = moderate exposure

4 = high exposure

Total N = 467

Group Means on The DV

1	2	3	4	Grand Mean
3.29	3.39	3.44	3.51	3.41

Basic One -Way Anova

Source	SS	df	MS	F	p
Exposure	3.36	3	1.12	7.14	.001
error	72.64	463	.16		

Coding Group Membership

	Group			
	1	2	3	4
Dummy				
D1	0	1	0	0
D2	0	0	1	0
D3	0	0	0	1
Effects				
E1	-1	1	0	0
E2	-1	0	1	0
E3	-1	0	0	1

Statistical Meaning of Codes

> The codes must indicate, as a group or set, unambiguously and non-redundantly each person's group.

> If accomplished, than the test of the basic null hypothesis that the groups are the same on the DV will be identical to the basic ANOVA results regardless of how group is measured or scaled.

Substantive Meaning of Codes

> The interpretation of the model parameters (partial regression coefficients) and the tests of those parameters will depend crucially on how the variables are coded.

> Specifically, the values and tests of the model parameters will concern specific comparisons among the groups.

> If you do not care how the groups differ (only that they do), then scaling matters not (but it seems that we should care?)

Illustration

```
>MODEL OPEN = CONSTANT + D1+D2+D3

Multiple R: 0.21   Squared multiple R: 0.04
     Effect    Coefficient     t   P(2 Tail)
       CONSTANT     3.29     92.77     0.00
       D1           0.10      1.93     0.05
       D2           0.15      2.93     0.00
       D3           0.22      4.51     0.00
             Analysis of Variance
Source              SS   df     MS   F-ratio       P
Regression        3.36    3   1.12    7.14        0.00
Residual         72.64  463   0.16
```

Illustration continued

```
>MODEL OPEN = CONSTANT + E1+E2+E3

Multiple R: 0.21   Squared multiple R: 0.04
     Effect      Coefficient    t    P(2 Tail)
      CONSTANT     3.41      184.76   0.00
      E1          -0.02       -0.57     0.57
      E2           0.03        1.05     0.29
      E3           0.10        3.38     0.00
           Analysis of Variance
Source            SS    df      MS   F-ratio       P
Regression      3.36     3    1.12      7.14      0.00
Residual       72.64   463    0.16

        Group Means on Openness

                          Group
                1        2        3        4
               3.29     3.39     3.44     3.51

        Grand Mean                 3.41
```

Illustration continued

$$Y' = C + b1(D1) + b2(D2) + b3(D3)$$

For Dummy codes

Group 1

$Y' = 3.29 + .10(0) + .15(0) + .22(0) = 3.29$

Group 2

$Y' = 3.29 + .10(1) + .15(0) + .22(0) = 3.39$

and so on.

Logistic Regression and Odds Ratios

```
>LOGIT

>MODEL CHD = CONSTANT+CESD+AGE+BMI+HDL+LDL+ACTIVITY

Log Likelihood:   -243.76691
```

Parameter	Estimate	S.E.	t-ratio	p-value
CONSTANT	-5.95882	1.41119	-4.22	0.00002
CESD	0.04354	0.01455	**2.99**	0.00277
AGE	0.05430	0.01179	4.61	0.00000
BMI	0.00078	0.02430	0.03	0.97432
HDL	-1.04931	0.40617	-2.58	0.00978
LDL	0.26651	0.13028	2.05	0.04079

Odds Ratio = e^{beta} (e = 2.71828…)

CESD	$e^{(.04354)}$ =	**1.04450**	Mean = 6.59; sd = 7.66
AGE		1.05580	
BMI		1.00078	
HDL		0.35018	
LDL		1.30540	

Logistic Regression and Odds Ratios

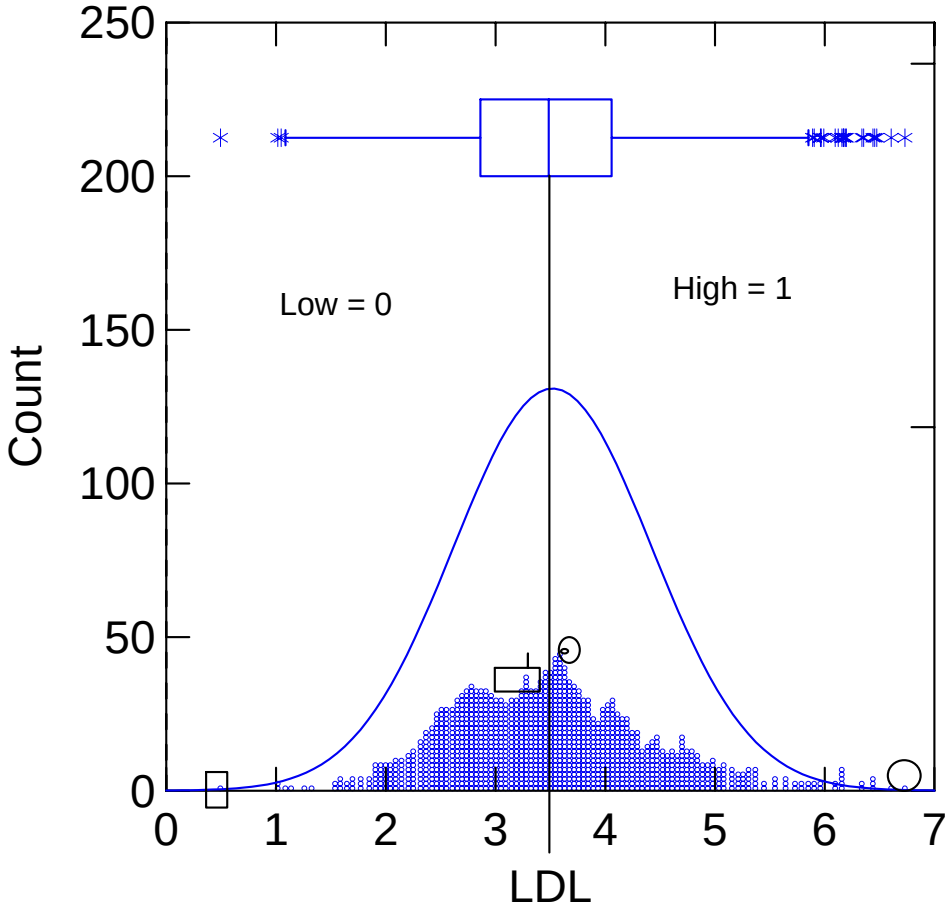
```
>LET CESD = CESD/100

Log Likelihood:   -243.76691
```

Parameter	Estimate	S.E.	t-ratio	p-value
CONSTANT	-5.95869	1.41119	-4.22	0.00002
CESD	4.35392	1.45506	**2.99**	0.00277
AGE	0.05430	0.01179	4.61	0.00000
BMI	0.00078	0.02430	0.03	0.97432
HDL	-1.04931	0.40617	-2.58	0.00978
LDL	0.26651	0.13028	2.05	0.04079

	Odds Ratio	
CESD	$e^{(4.35392)}$ = **77.78303**(!)	M = .0659; S = .0766
AGE	1.05580	
BMI	1.00078	
HDL	0.35018	
LDL	1.30540	

The Median Split



Importance of Sample

Reliability and Validity are NOT PURE properties of measures.

> "*Reliability coefficients for X are reported to range from .70 to .85. According to the manual (ref) X has been found to be a valid measure of A.*" ---Anon

A measure's psychometric characteristics are necessarily a function of the sample on which they are used and the purpose for which they are used.

-----Evaluate the measure on your data (most easily done with coefficient alpha)

-----This means you must enter the items, not just the total score in your data set.

-----Do not assume that a measure that is acceptable in one sample will be acceptable in a different one. Check it!

A personally distressing example

Juror Bias Scale

Kassin, S. M., & Wrightsman, L. S. (1983). The construction and validation of a juror bias scale. *Journal of Reserch in Personality, 17,*423-442.

17 items; split half reliability = .81 in a sample of 221 college students in a mock jury setting

Clark and Chaplin

Administered the juror bias scale to 388 individuals who were serving in a jury pool in Alabama to see, among other things, if bias was related to selection.

But our estimate of split-half reliability for this scale in this sample is .59. (!)

We are now writing a far different, and less exciting manuscript.