

Measurement and RCTs: Relevance and Developments

William F. Chaplin, Ph.D.
Department of Psychology
St. Johns University
New York

Why Is Measurement Important for Our Research?

- >Type II error

 - (power)

- >Type III error

 - (doing the wrong study)

- >Interpreting model parameters

 - (translating results into treatment and policy)

 - (that is, clinical and substantive in addition to statistical significance)

Reliability and Type II error

Measurement unreliability attenuates the obtained relations among variables.

$$\rho_{xy} \sqrt{r_{xx}} \sqrt{r_{yy}} = \mathbf{r}_{xy}$$

The diagram illustrates the relationship between the true correlation and the observed correlation. The equation $\rho_{xy} \sqrt{r_{xx}} \sqrt{r_{yy}} = \mathbf{r}_{xy}$ is shown. Arrows point from labels to the terms in the equation: 'True relation' points to ρ_{xy} , 'Reliability of x' points to $\sqrt{r_{xx}}$, 'Reliability of y' points to $\sqrt{r_{yy}}$, and 'Observed relation' points to \mathbf{r}_{xy} .

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

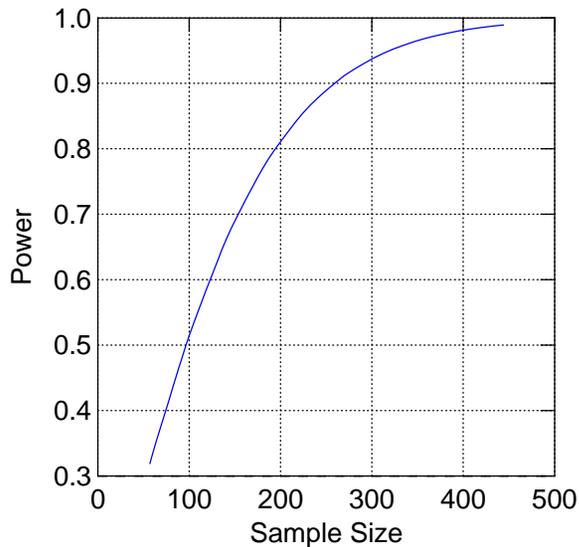
If the reliability of the measure of clinical improvement is .80 then the expected observed treatment–outcome relation is .178; if the reliability is .70 then the observed relation will be .167 and so on.

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

Implications for Statistical Power

Power set at .80

Power Curve (Alpha = 0.050)

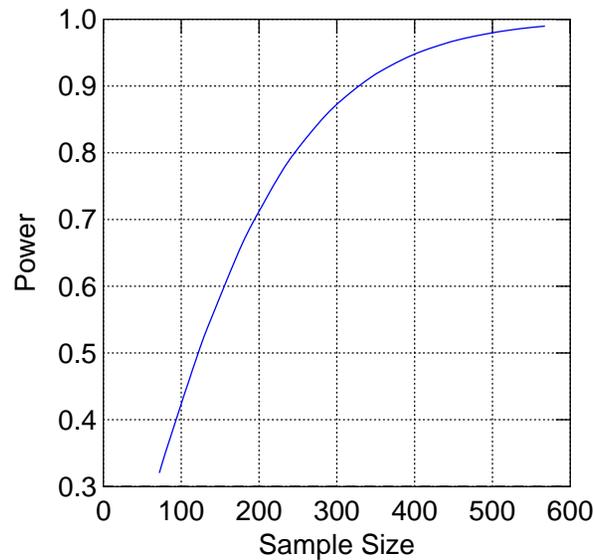


$r = .20$

Needed N = 194

(actual study N = 1,850)

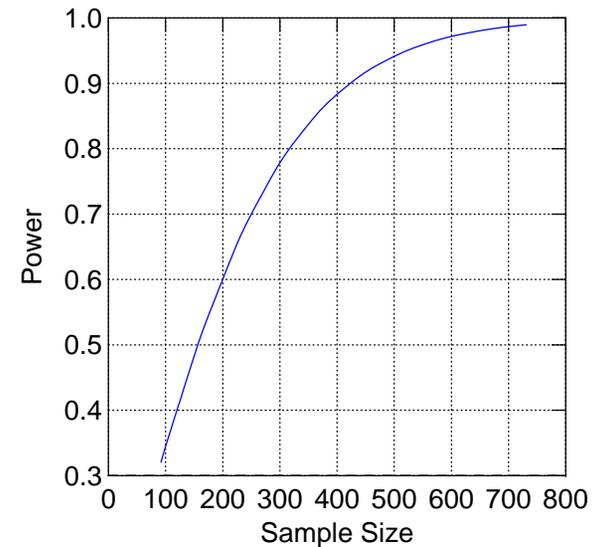
Power Curve (Alpha = 0.050)



$r = .178$

246

Power Curve (Alpha = 0.050)



$r = .157$

316

Correction for Attenuation

$$\rho_{xy} = r_{xy} / (\sqrt{r_{xx}} \sqrt{r_{yy}})$$

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 ($I_{\text{new}} / I_{\text{old}}$)

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

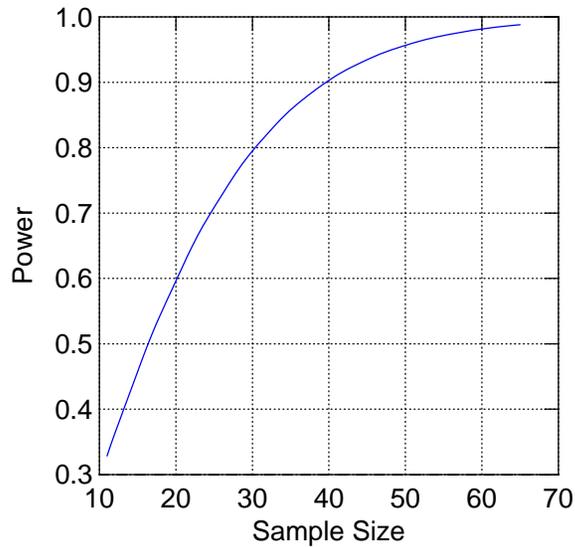
The hypothesized relation between aggression and narcissism (based on perfect measures) is .49.

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

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

Effect on Power Analysis for power of .80

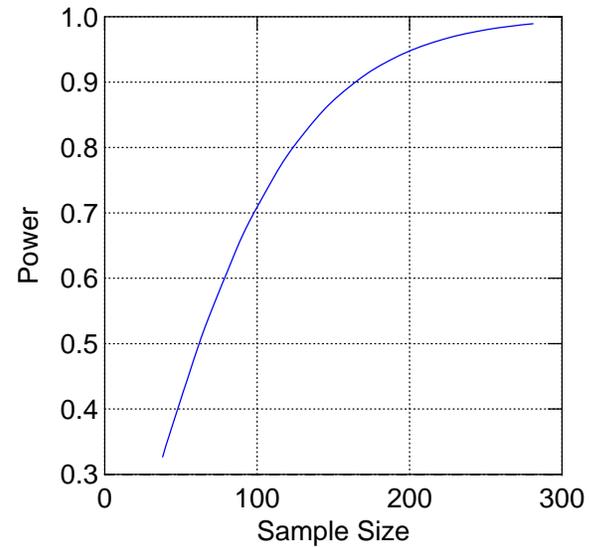
Power Curve (Alpha = 0.050)



$$r = .49$$

$$N = 31$$

Power Curve (Alpha = 0.050)



$$r = .25$$

$$N = 124$$

The Effect of Unreliability of Categorization

Adapted from Horton, N. J., & Shapiro, E. C. (2005). Statistical Sleuthing During Epidemics: Maternal Influenza and Schizophrenia. *Chance*, 18, 11-18.

Is maternal influenza linked to schizophrenia in offspring?

“In many of these studies maternal influenza was defined as ‘exposure to an influenza epidemic’.... This indirect method likely **mismeasures** exposure for some subjects.” (p. 15)

Such measurement error is illustrated in a hypothetical study of 10,000 subjects, 1% of whom have schizophrenia and 10% with **true** maternal exposure. Based on previous research the relative risk of maternal exposure to influenza on offspring schizophrenia is 1.588

1) No error of exposure assessment

Based on the relative risk of 1.588 we get

		Schizophrenia		
		Yes	No	
Truly Exposed	Yes	15	985	1000
	No	85	8915	9000
		100	9900	10,000

2) Exposure Classification 90% accurate

		Schizophrenia		
		Yes	No	
Observed Exposure	Yes	22	1778	1800
	No	78	8122	8200
		100	9900	10,000

$.9(15) + .1(85)$
 $.9(985) + .1(8915)$

$$2) \text{ RR} = (22/1800)/(78/8200) = .0122/.0095 = 1.28$$

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 how 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

Total N = 467

2 = low exposure

3 = moderate exposure

4 = high exposure

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		

ANOVA models are limited to categorical variables.

Regression models are more general as they allow both categorical (experimental) and quantitative (individual difference) variables in the same model. Especially useful for evaluating moderator and mediator models. However, to do this we must represent (*measure*) the categories appropriately.

From a statistical perspective there are essentially an infinite number of ways we can do this. The requirement is simply that the coding communicate

Unambiguously,

Completely, and

Non-redundantly

to what group a participant has been assigned

However, there is a small set of these codes that also make the parameter estimates associated with the set of coded variables ***Meaningful.***

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, then 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

$$Y' = C + b_1(D_1) + b_2(D_2) + b_3(D_3)$$

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.

Centering Quantitative Variables

A centered variable is simply one that is expressed as deviation scores.

It is “between” a raw score and a standard (z) score in that it standardizes the mean at 0, but leaves the units of measurement intact.

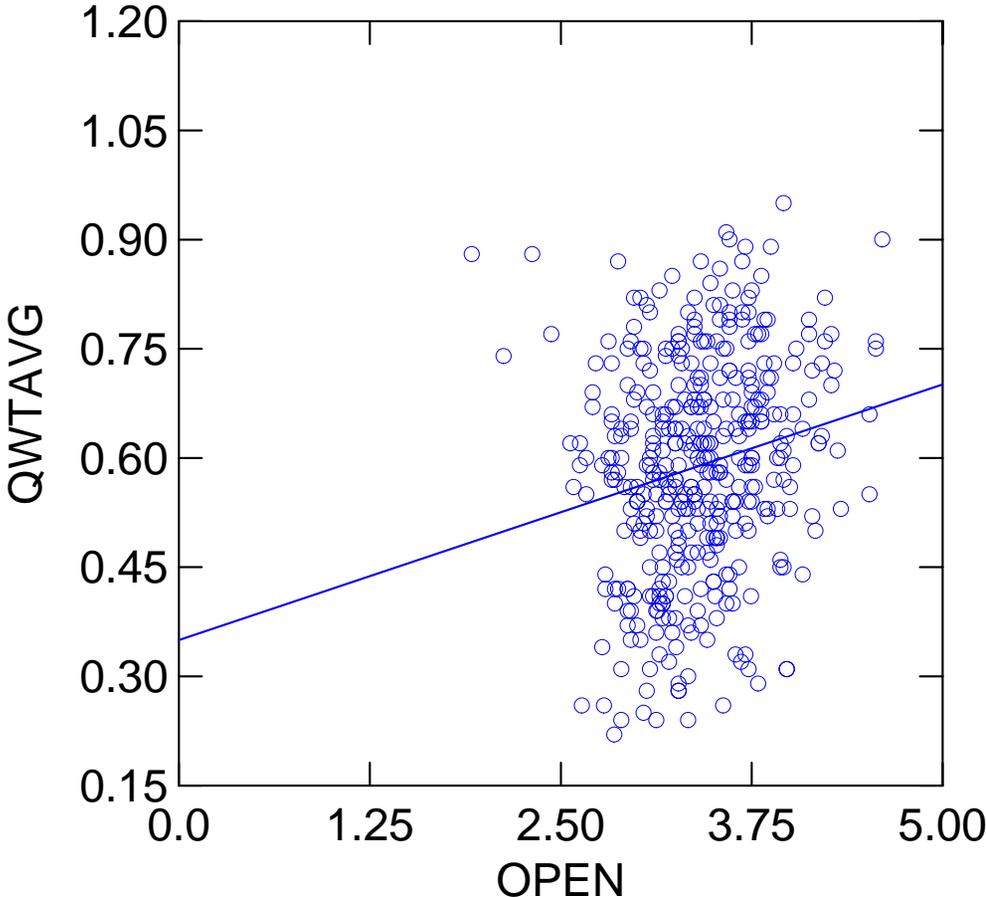
Centered Variable = Original Variable – Sample Mean.

Why do this.....

Dep Var: QWTAVG N: 408 Multiple R: 0.191 Squared multiple R: 0.036

Adjusted squared multiple R: 0.034 Standard error of estimate: 0.146

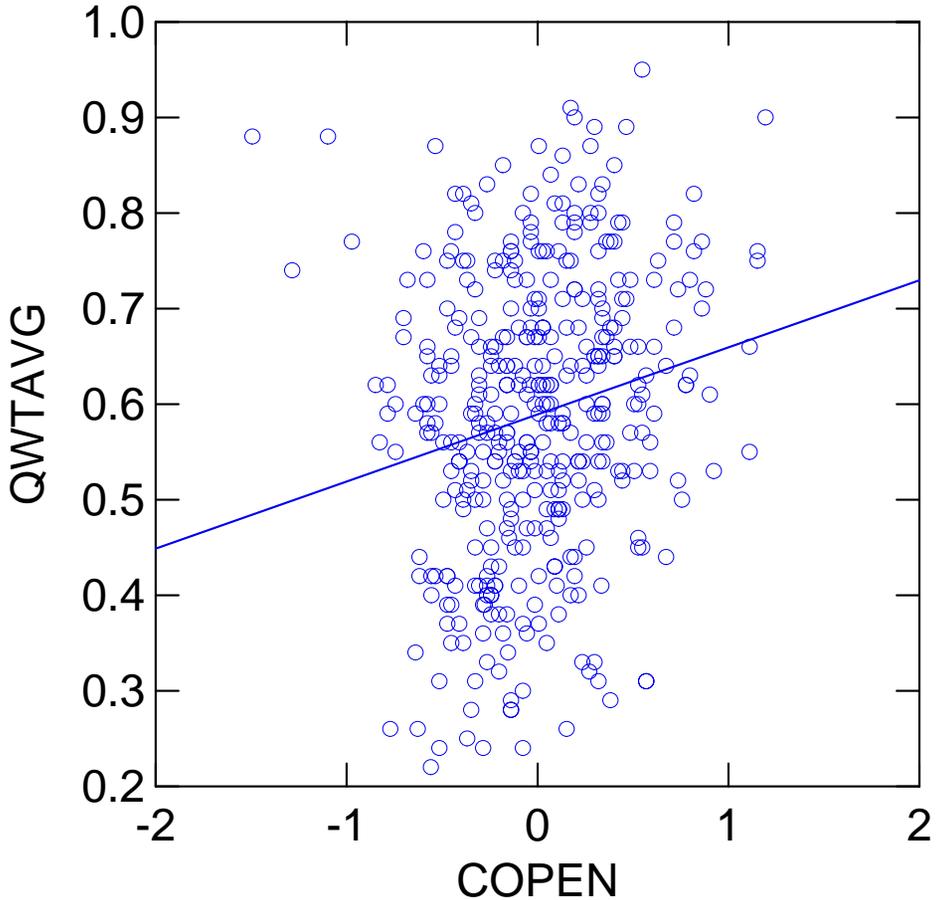
Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	0.350	0.061	0.0	.	5.699	0.000
OPEN	0.070	0.018	0.191	1.000	3.921	0.000



Dep Var: QWTAVG N: 408 Multiple R: 0.191 Squared multiple R: 0.036

Adjusted squared multiple R: 0.034 Standard error of estimate: 0.146

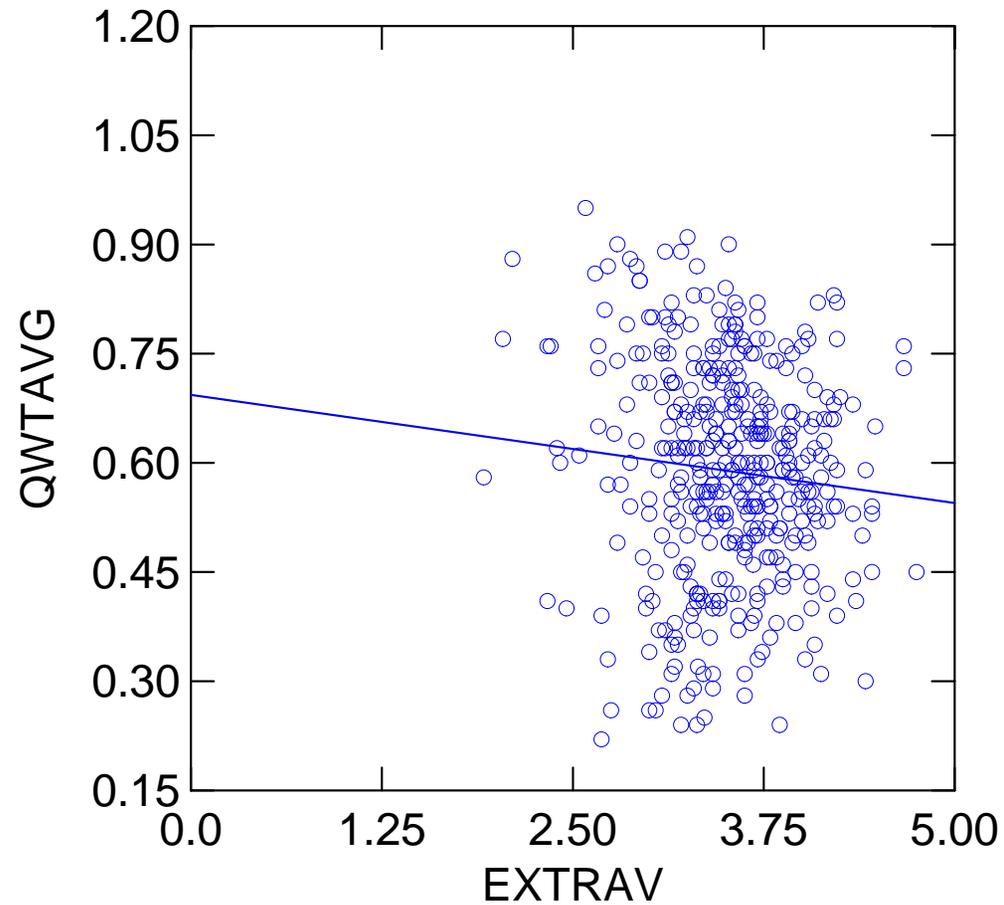
Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	0.589	0.007	0.0	.	81.347	0.000
COPEN	0.070	0.018	0.191	1.000	3.921	0.000



Dep Var: QWTAVG N: 408 Multiple R: 0.089 Squared multiple R: 0.008

Adjusted squared multiple R: 0.006 Standard error of estimate: 0.148

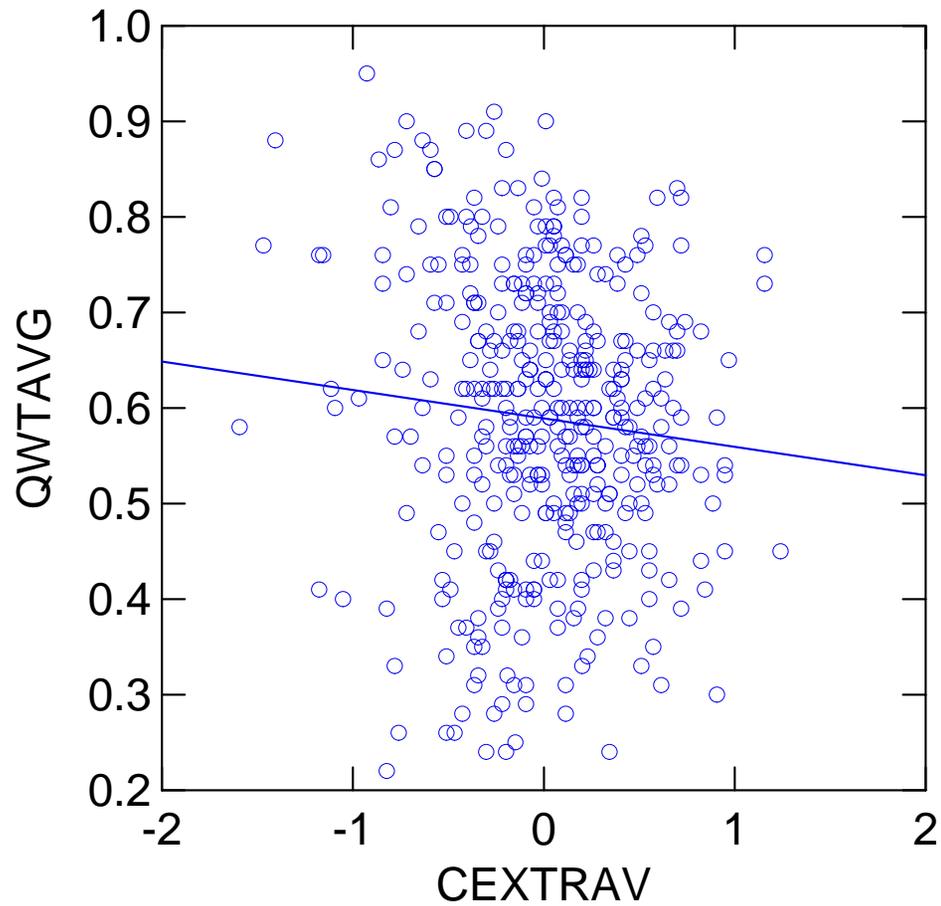
Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	0.693	0.058	0.0	.	11.900	0.000
EXTRAV	-0.030	0.016	-0.089	1.000	-1.809	0.071



Dep Var: QWTAVG N: 408 Multiple R: 0.089 Squared multiple R: 0.008

Adjusted squared multiple R: 0.006 Standard error of estimate: 0.148

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	0.589	0.007	0.0	.	80.137	0.000
CEXTRAV	-0.030	0.016	-0.089	1.000	-1.809	0.071



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
```

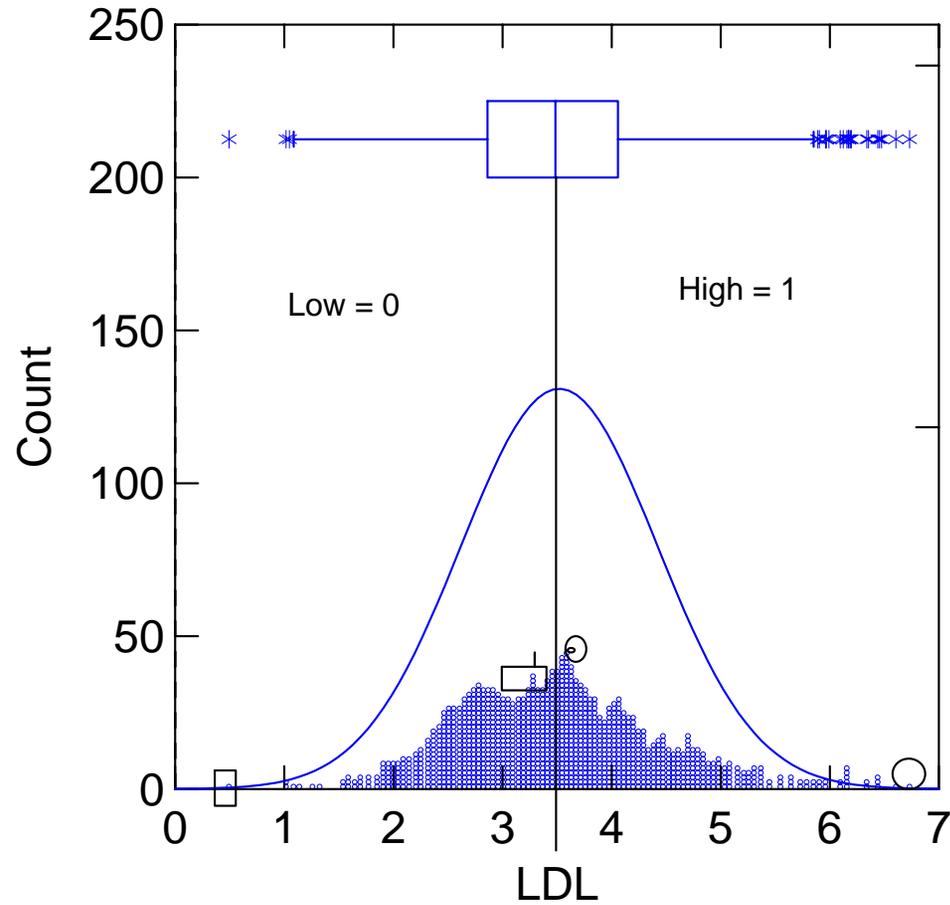
```
Log Likelihood: -243.76691
```

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

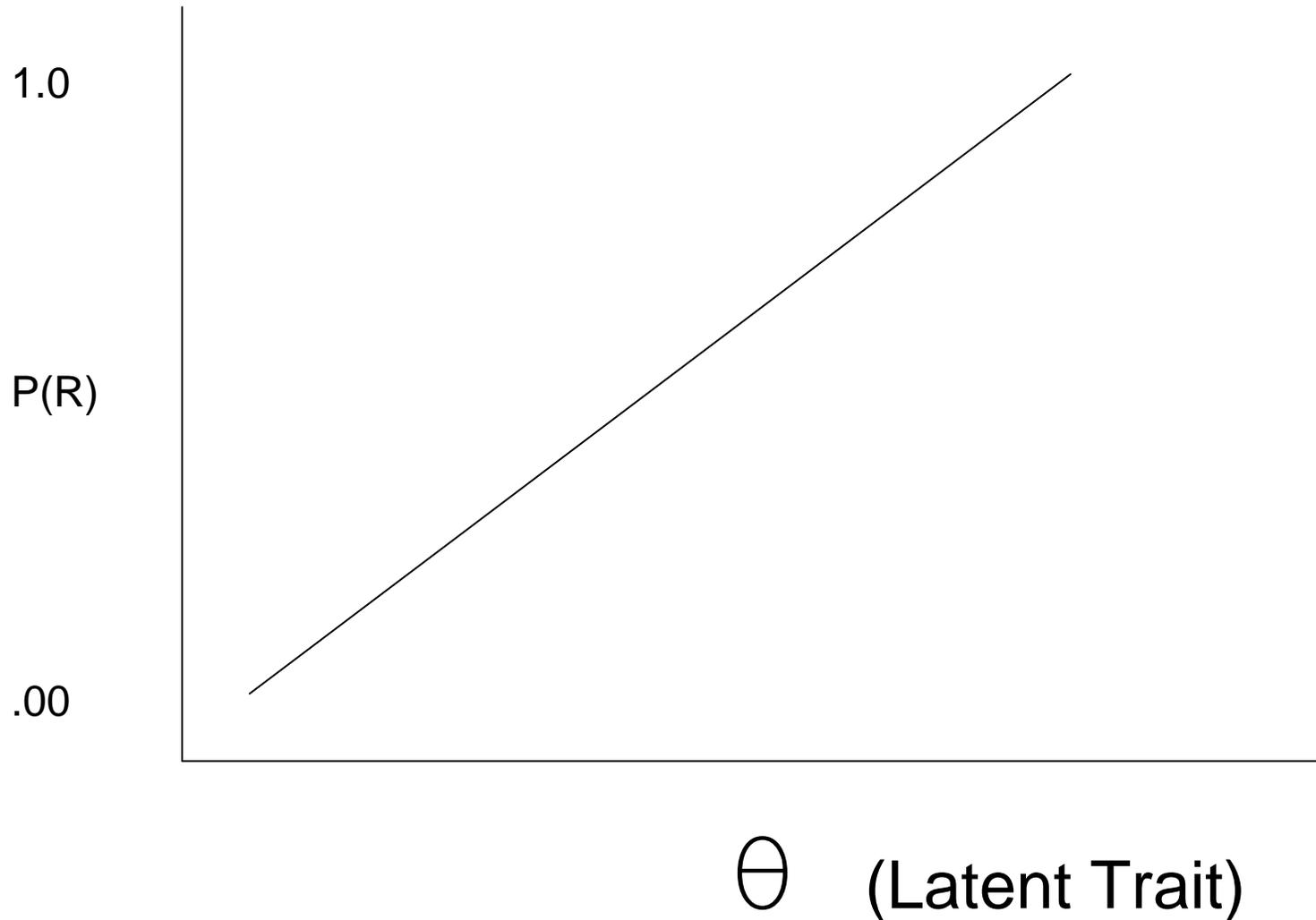


Item Response Theory (IRT)

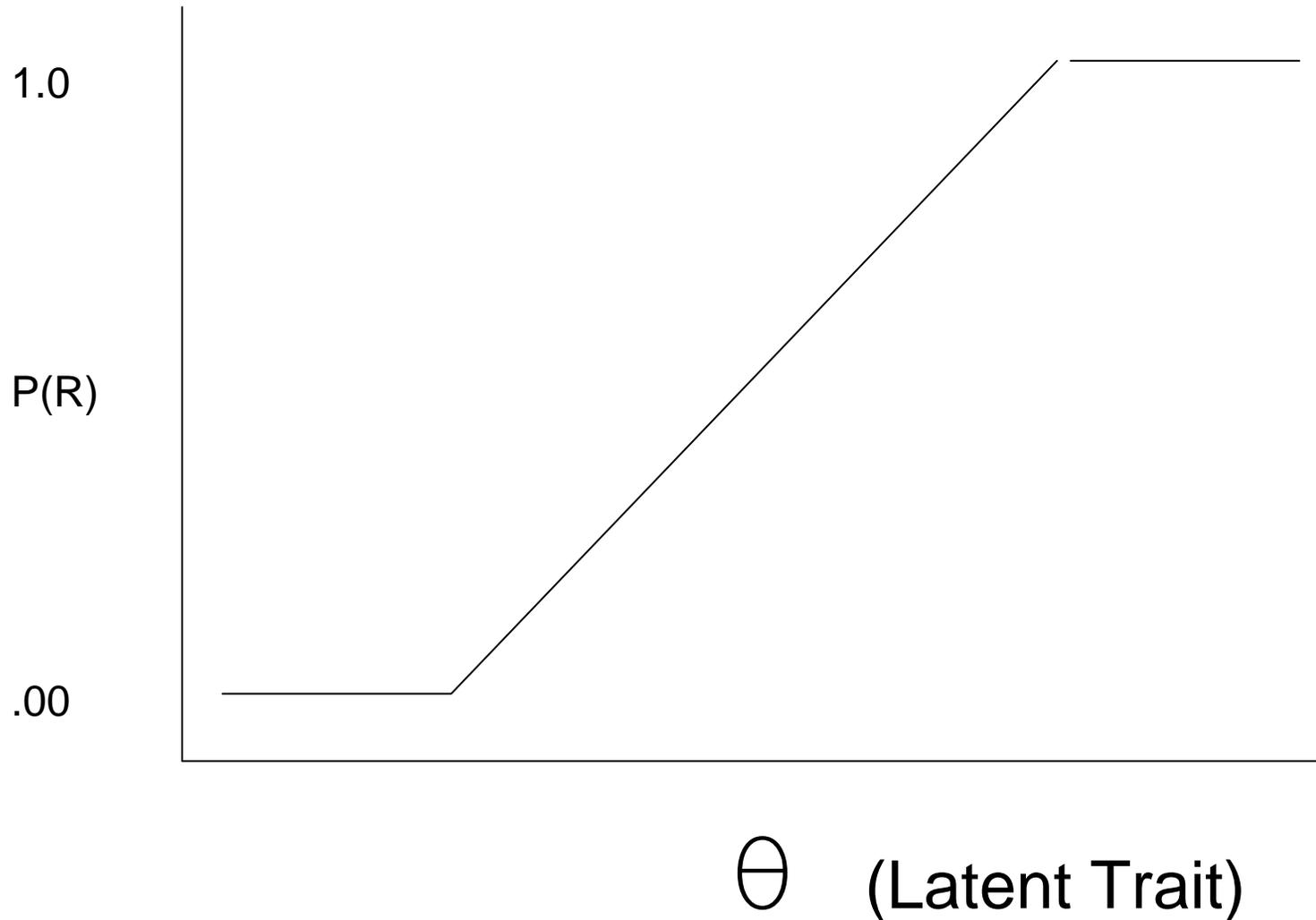
All measurement ultimately is based on our theories of how a person's observed response to an item is related to the underlying characteristic that we are trying to measure.

In everyday measurement these theories of item response are often not specified. IRT seeks to make explicit and specific what has historically been implicit and vague.

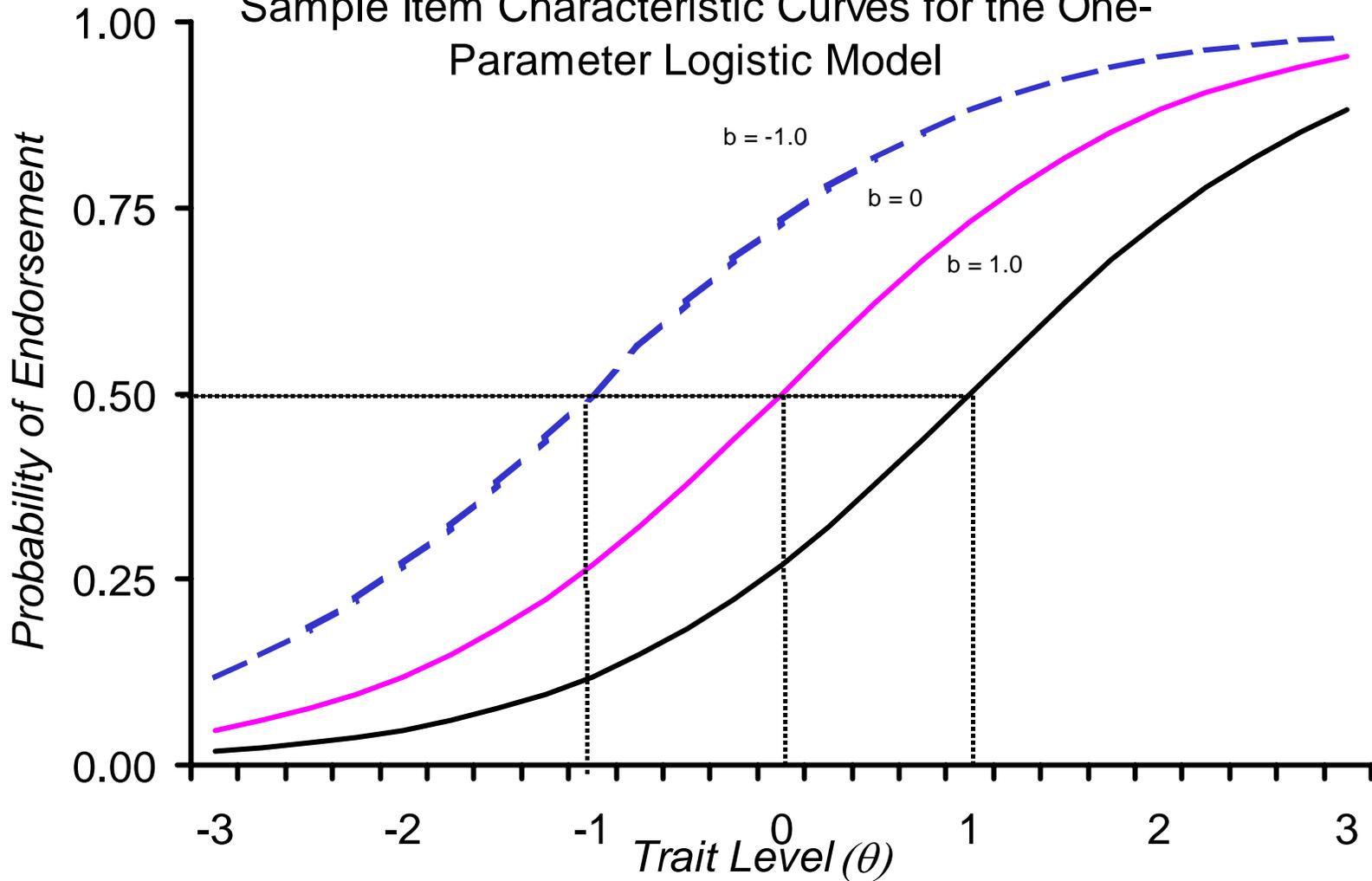
Item Characteristic Curve (ICC)



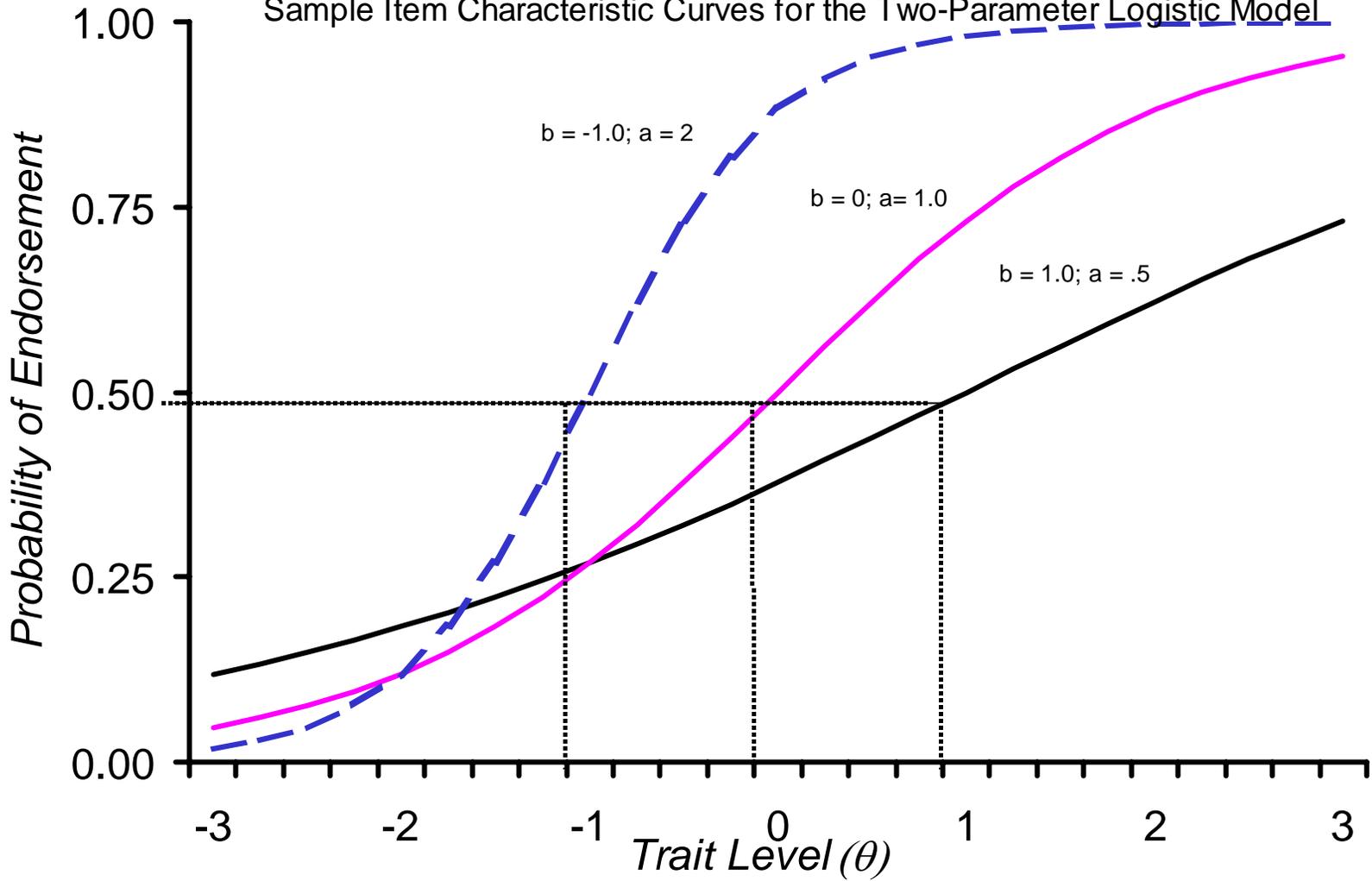
Item Characteristic Curve (ICC)



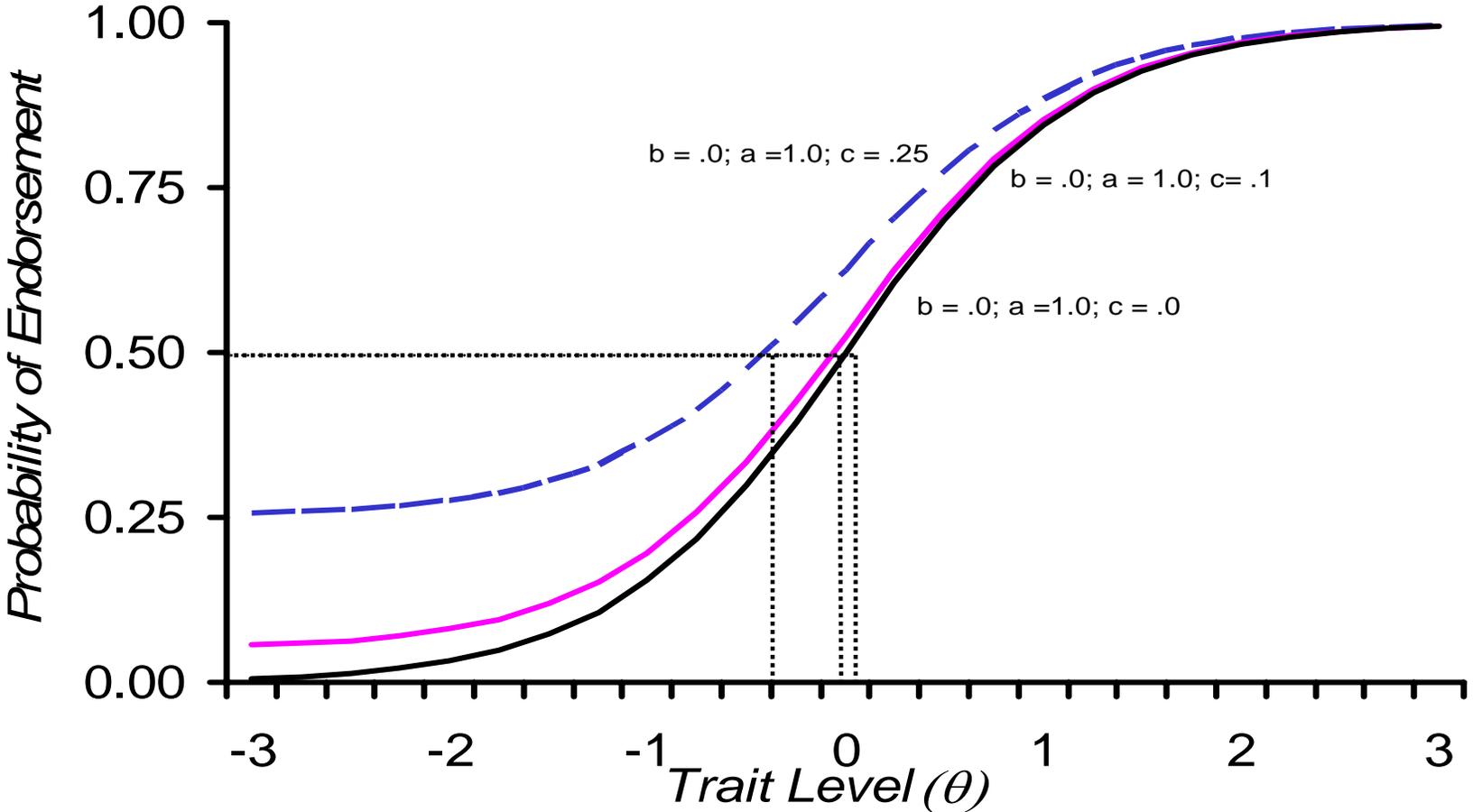
Sample Item Characteristic Curves for the One-Parameter Logistic Model



Sample Item Characteristic Curves for the Two-Parameter Logistic Model



Sample Item Characteristic Curves for the Three-Parameter Logistic Model



1. Three-Parameter Logistic Model

$$P(X_{is} = 1 | \theta_s, b_i, a_i, c_i) = c_i + (1 - c_i) \frac{\exp[a_i(\theta_s - b_i)]}{1 + \exp[a_i(\theta_s - b_i)]}$$

X_{is} = response of person s to item i (0 or 1)

θ_s = trait level for person s

b_i = difficulty or threshold of item i

a_i = discrimination of item i

c_i = the lower asymptote of item i

\exp = the natural log base (2.718)

Estimation of parameters proceeds by iterative maximum likelihood procedures.

IRT provides a very strong model that can lead to precise estimates of the person's characteristic and to the item characteristics.

Used to develop measures as well as scale people.

Identifies DIF when applied to identifiable groups.

After identifying DIF for some items those items can be eliminated from scale, or estimates of construct can be based on different measurement models.

DIF can be an interesting substantive finding in itself.

Readable Introduction:

Hambleton, R. K., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage

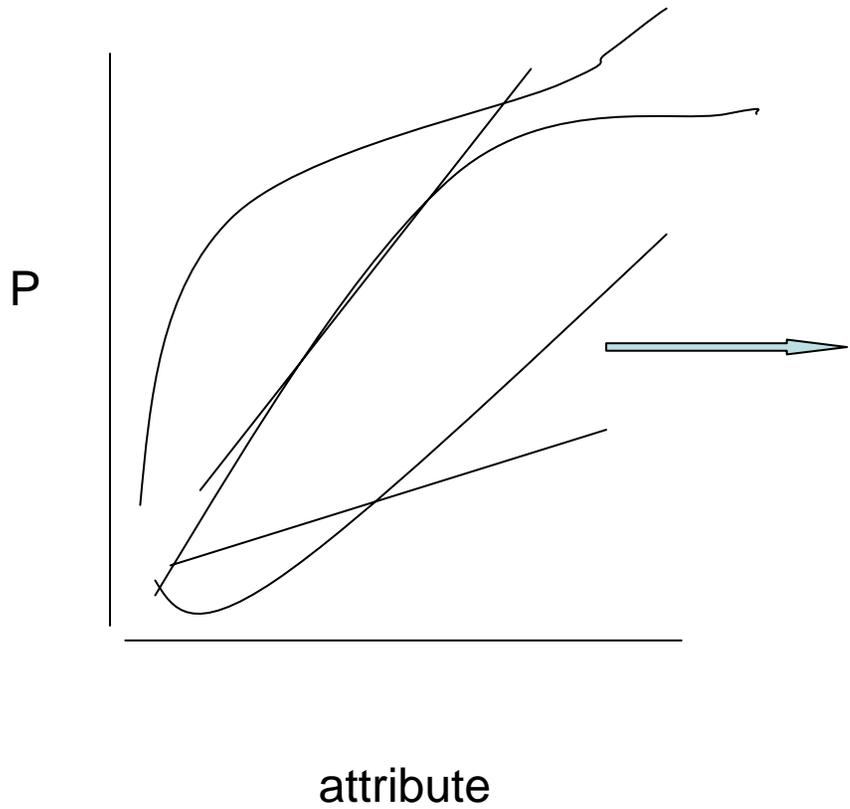
Most common scoring of a scale is to sum (or average) the items.

This is based on a powerful, but vague IRT called the “General Linear Scaling Model”

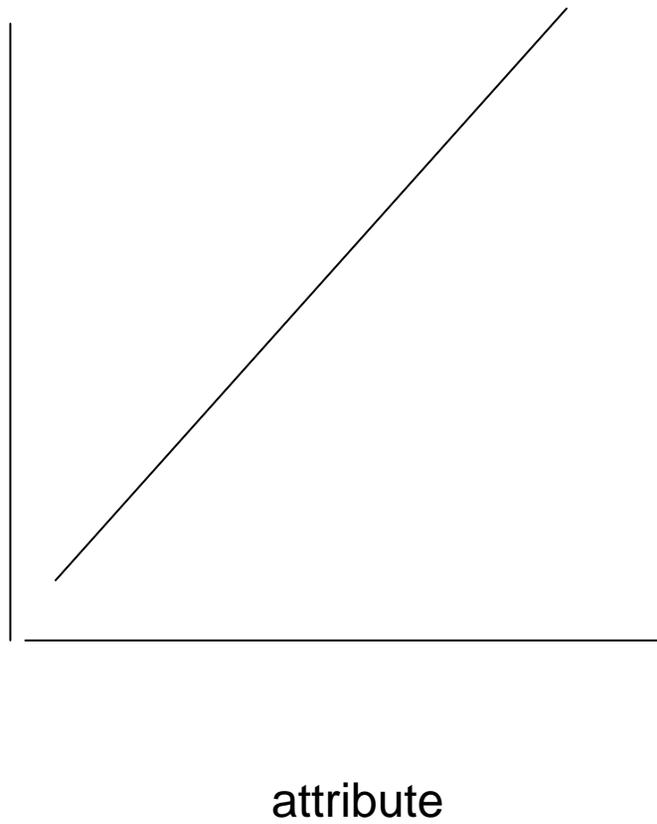
Assumptions:

Items are unidimensional

ICC's are monotonic



s
c
o
r
e



Confirmatory Factor Analysis (CFA)

A form of structural equation modeling (SEM) often called the “measurement model” component of SEM.

CFA allows you to test that your items have particular structural properties such as unidimensionality or a particular subscale structure

$$\left[\begin{array}{c} \phantom{r_{ij}} \\ \phantom{r_{ij}} \\ \phantom{r_{ij}} \\ \phantom{r_{ij}} \end{array} \right] r_{ij} - \left[\begin{array}{c} \phantom{r_{ij}} \\ \phantom{r_{ij}} \\ \phantom{r_{ij}} \\ \phantom{r_{ij}} \end{array} \right] \text{Est } r_{ij}$$

$$= \left[\begin{array}{c} \phantom{r_{ij}} \\ \phantom{r_{ij}} \\ \phantom{r_{ij}} \\ \phantom{r_{ij}} \end{array} \right] \text{Res } r_{ij}$$

Researchers who apply CFA techniques “do not seem adequately sensitive to the fundamental reality that there is no true model....and that the best one can hope for is to identify a parsimonious, substantively meaningful model that fits observed data adequately well.” (p. 213)

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201-226.

Example—MH LCS

MH LCS developed by Wallston et.al. to assess three dimensions of health locus of control beliefs Internal (I), and two External Dimensions, Powerful Others (PO) and Chance (C)

In 1996 a revised MH LCS was developed that added a third External Dimension (God (G))

We used CFA to evaluate the structure of the MHLCS to see if its structure was consistent with the hypothesized dimensions

The revised MHLCS has 24 items; 6 items on each of the four subscales

Chaplin, W. F., Davidson, K. W., Sparrow, V. M., Stuhr, J., Van Roosemalen, E., & Wallston, K. A. (2001) A psychometric and structural evaluation of the expanded Multidimensional Health Locus of Control Scale with a diverse sample of Caucasian/European, Native, and Black Canadian women. British Journal of Health Psychology, 6, 447-455.

Correlations Among the Four Health Locus of Control Subscales

Subscale	1	2	3	4
1. Internal	---	.10	.14	.14
2. Chance		---	.54	.50
3. God			---	.43
4. Powerful Others				---

Note. N = 371. Correlations larger than .11 are significant at the .05 level, two-tailed test.

Summary of the Goodness of Fit Indices for Different Structural Models of the Health Locus of Control Items

<u>Model</u>	<u>Degrees of Freedom</u>	<u>Chi-Square</u>	<u>S-B Scaled Chi-Square</u>	<u>CFI</u>	<u>Robust CFI</u>	<u>Standardized RMSR</u>
Independence	276	2679.7	-----	----	----	----
1 Factor	252	963.4	763.4	.704	.722	.089
4 Independent Factors	252	763.3	625.3	.787	.797	.156
4 Correlated Factors	246	490.4	399.6	.898	.916	.067
2 Independent Factors (Internal vs External)	252	764.3	610.8	.787	.805	.084
2 Correlated Factors	251	762.0	609.6	.787	.805	.082
4 Factors with three External Factors Correlated	249	492.8	401.2	.899	.917	.069

Note. N = 371. S-B = Sattora-Bentler; CFI = Comparative Fit Index; RMSR = Root Mean Squared Residual

Standardized Estimates of the Path Coefficients for the Four Factor Model with the
"External" Factors Correlated for the HLCS

Item	Factor				Error
	Internal	Chance	God	Powerful Others	
I1	.36				.93
I2	.35				.94
I3	.14*				.99
I4	.58				.81
I5	.58				.81
I6	.73				.68
C1		.46			.89
C2		.41			.91
C3		.53			.85
C4		.55			.84
C5		.47			.88
C6		.60			.80
G1			.73		.68
G2			.49		.87
G3			.78		.63
G4			.78		.62
G5			.84		.54
G6			.83		.56
P1				.47	.89
P2				.50	.87
P3				.10*	.99
P4				.61	.79
P5				.60	.80
P6				.72	.70

Correlations Among the Factors

	Chance	God	Powerful Others
Chance	----		
God	.71	----	
Powerful Others	.77	.56	----

Note. N = 371. I = Internal, C = Chance, G = God, P = Powerful Others. Items are grouped by factor. The number of the item corresponds to the order it appears on the MHLCS. All paths are

Significant except those marked with a *

Coefficient Alpha and Range of Item Total Correlations

for the MHLCS Scales

<u>Scale</u>	<u>Coefficient Alpha</u>	<u>Range of Corrected Item-Total Correlations</u>
Internal	.60	.14 - .45
Chance	.68	.35 - .47
God	.88	.47 - .77
Powerful Others	.65	.10 - .49