Reliability Theory
Classical Model (Model T)

Problem: To what extent does a measure represent a construct? I.e., what percentage of variance is construct variance? What is the correlation between observed measure and latent construct? For any particular observed score, what is the most likely latent score?

Classical Model:

Observed score = Latent score + Error

Observed = True + Error

$X = T + E$ or $x = t + e$

Define True score as the expectation of observed score.
(Note that this is not the same as Platonic Truth).
Then Error is uncorrelated with True (latent) score since the mean error for any X is zero.

Variance of X = $V_x = V_{(t+e)} = V_t + V_e + 2C_{te} = V_t + V_e$ and $V_t/V_x = V_t/(V_t + V_e)$ = percentage of test variance which is true score variance.

Covariance between observed and latent score is

$$C_{xt} = \frac{\sum(xt)}{N} = \frac{\sum[(t+e)*t]}{N} = \frac{\sum(t^2)}{N} = V_t$$

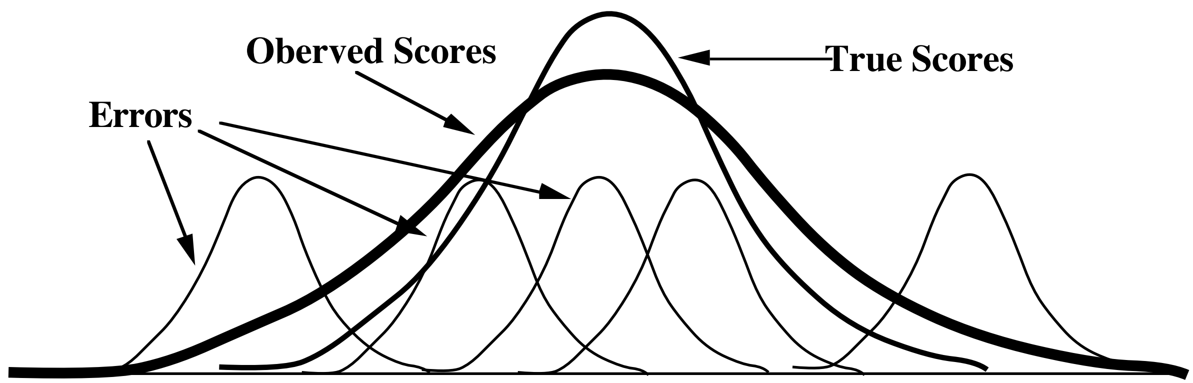
Correlation (rho) between x and t =

$$r_{xt} = \frac{C_{xt}}{\sqrt{V_x * V_t}} \quad ==>$$

$$r_{xt} = \frac{V_t}{(S_x * S_t)} = \frac{S_t}{S_x} \quad ==> \quad r_{xt}^2 = \frac{V_t}{V_x}$$

Given x, the most likely true score can be found by regression: $z_t = r_{xt} * z_x$

The relationship between True, Error, and Observed Scores. Observed = True + Error ==> Obs erved score variance > True Score variance.



To predict true score from an observed score:
$z_{t(\text{predicted})} = r_{xt} * z_x$ ==> $t/S_t = r_{xt} * x/S_x$ ==>
$t = r_{xt} * x * S_t/S_x$ =. But $r_{xt} = S_t/S_x$ and $r_{xx} = V_t/V_x$

==>

$$t_{(\text{predicted})} = r_{xx} * x.$$

$C_{xe} = \sum(x*e)/N = \sum[(t+e)*e]/N = V_e$

$$r_{xe} = \frac{C_{xe}}{\sqrt{V_x \square V_e}} = \frac{V_e}{(S_x * S_e)} = \frac{S_e}{S_x} = \sqrt{\frac{V_e}{V_x}}$$

$V_e = V_x - V_t = V_x * (1 - r_{xx})$ ==> $r_{xe} = \sqrt{1 - r_{xx}}$

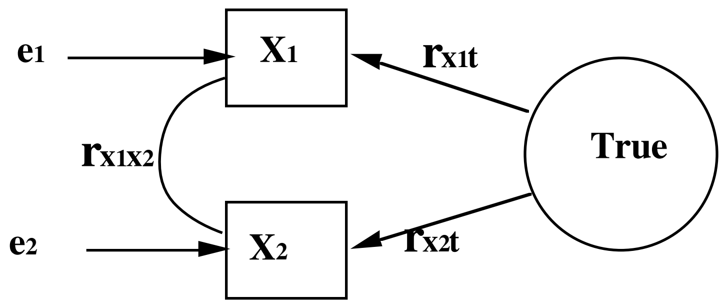
Observed score is correlated with error.

Obs erved score variance =true score variance + error score variance

The problem remains, however, how do we find V_t or V_{xt}?

Classical theory of parallel tests

Consider two tests (X_1 and X_2) which both measure the same construct (T). Assume that for every individual that $t_1 = t_2 = t$.



$$x_1 = t + e_1 \qquad x_2 = t + e_2$$

Even if $e_1 \neq e_2$, we can assume that $V_{e_1} = V_{e_2}$ Then:
$V_{x_1} = V_t + V_{e_1} = V_t + V_{e_2} = V_{x_2} = V_x$

$$C_{x_1x_2} = \frac{\sum(x_1 * x_2)}{N} = \frac{\sum[(t+e_1)(t+e_2)]}{N} \implies$$

$$C_{x_1x_2} = V_t + C_{te_1} + C_{te_2} + C_{e_1e_2} = V_t$$

$$r_{x_1x_2} = \frac{C_{x_1x_2}}{\sqrt{V_{x_1} * V_{x_2}}} = \frac{C_{x_1x_2}}{V_x} = \frac{V_t}{V_x}$$

$$r_{x_1x_2} = \frac{V_t}{V_x} = r_{xt}^2$$

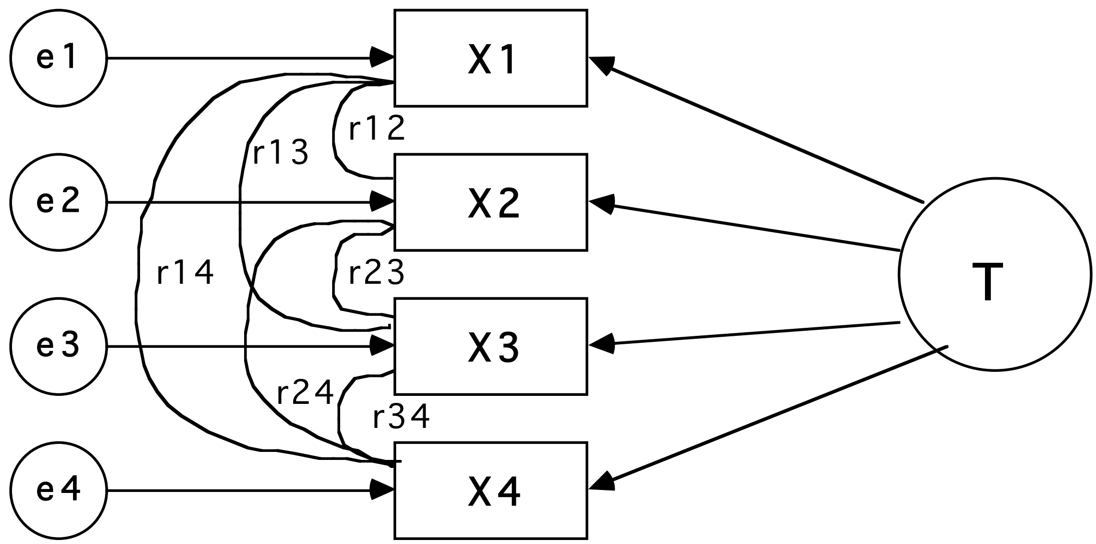
The reliability is the correlation between two parallel tests and is equal to the squared correlation of the test with the construct.

$r_{xx} = \frac{V_t}{V_x}$ = percent of test variance which is construct variance.

$r_{xt} = \sqrt{r_{xx}}$ ==> the validity of a test is bounded by the square root of the reliability.

How do we tell if one of the two "parallel" tests is not as good as the other? That is, what if the two tests are not parallel?

Congeneric Measurement Theory



This matrix will have the following covariances:

	x_1	x_2	x_3	x_4
x_1	V_{x_1}			
x_2	$C_{x_1x_2}$	V_{x_2}		
x_3	$C_{x_1x_3}$	$C_{x_2x_3}$	V_{x_3}	
x_4	$C_{x_1x_4}$	$C_{x_2x_4}$	$C_{x_3x_4}$	V_{x_4}

These covariances reflect the following parameters:

	x_1	x_2	x_3	x_4
x_1	$V_t+V_{e_1}$			
x_2	$C_{x_1t}C_{x_2t}V_t$	$V_t+V_{e_2}$		
x_3	$C_{x_1t}C_{x_3t}V_t$	$C_{x_2t}C_{x_3t}V_t$	$V_t+V_{e_3}$	
x_4	$C_{x_1t}C_{x_4t}V_t$	$C_{x_2t}C_{x_4t}V_t$	$C_{x_3t}C_{x_4t}V_t$	$V_t+V_{e_4}$

We need to estimate the following parameters:
V_t, V_{e1}, V_{e2}, V_{e3}, V_{e4}, C_{x1t}, C_{x2t}, C_{x3t}, C_{x4t}

Parallel tests assume $V_{e1}=V_{e2} = V_{e3} = V_{e4}$, and $C_{x1t}= C_{x2t} = C_{x3t} = C_{x4t}$ and only need two tests.

Tau equivalent tests assume: $C_{x1t}=C_{x2t}= C_{x3t} =C_{x4t}$ and need at least three tests to estimate parameters.

Congeneric tests allow all parameters to vary but require at least four tests to estimate parameters.

Domain Sampling Theory-1

Consider a domain (D) of k items relevant to a construct. (E.g., English vocabulary items, expresions of impulsivity). Let D_i represent the number of items in D which the ith subject can pass (or endorse in the keyed direction) given all D items. Call this the domain score for subject i. What is the correlation of scores on an $item_j$ with domain scores?

$$C_{jd} = V_j + \sum_{l=1}^{k} C_{lj} = V_j + (k-1)*(\text{average covariance of } j)$$

$$\text{Domain Variance} = \sum_{l=1}^{k} V_l + \sum_{j \neq l}^{k} C_{lj} = \sum(\text{variances}) + \sum(\text{covariances})$$

V_d = k*(average variance) + k*(k-1) * (average covariance)
Let V_a = average variance and C_a = average covariance
then $V_d = k(V_a + (k-1)C_a)$.
Assume that $V_j = V_a$ and that $C_{jl} = C_a$.

$$r_{jd} = \frac{C_{jd}}{\sqrt{V_j * V_d}} = \frac{V_a + (k-1)*C_a}{\sqrt{V_a * k(V_a + (k-1)C_a)}}$$

$$r_{jd}^2 = \frac{(V_a+(k-1)*C_a)*(V_a+(k-1)*C_a)}{V_a*k*(V_a+(k-1)C_a)}$$

Now, find the limit of r_{jd}^2 as k becomes large:

$$\lim_{k \to \infty} r_{jd}^2 = \frac{C_a}{V_a} = \text{average covariance/average variance}$$

i.e., the amount of domain variance in an item (the squared correlation of the item with the domain) is the averge intercorrelation in the domain.

Domain Sampling Theory-2

What is the correlation of a test of n items with the domain score?

$$\text{Domain Variance} = \sum_{l=1}^{k} V_l + \sum_{j \neq l}^{k} C_{lj} = \sum(\text{variances}) + \sum(\text{covariances})$$

Let V_a = average variance and C_a = average covariance then $V_d = k(V_a + (k-1)C_a)$, $C_{nd} = n*V_a + n*(k-1)C_a$

V_n = variance of an n-item test = $\sum V_j + \sum C_{jl}$ =
$V_n = n*V_a + n*(n-1)*C_a$

$$r_{nd} = \frac{C_{nd}}{\sqrt{V_n * V_d}} \quad ==> \quad r_{nd}^2 = \frac{C_{nd}^2}{V_n * V_d} \quad ==>$$

$$r_{nd}^2 = \frac{\{n*V_a + n*(k-1)C_a\}*\{n*V_a + n*(k-1)C_a\}}{\{n*V_a + n*(n-1)*C_a\}*\{k(V_a + (k-1)C_a)\}}$$

$$r_{nd}^2 = \frac{\{V_a + (k-1)C_a\}*\{n*V_a + n*(k-1)C_a\}}{\{V_a + (n-1)*C_a\}*\{k(V_a + (k-1)C_a)\}} \quad ==>$$

$$r_{nd}^2 = \frac{\{n*V_a + n*(k-1)C_a\}}{\{V_a + (n-1)*C_a\}*\{k\}}$$

$$\lim \text{ as } k \to \infty \text{ of } r_{nd}^2 = \frac{n*C_a}{V_a + (n-1)C_a}$$

i.e., the amount of domain variance in a n-item test (the squared correlation of the test with the domain) is a function of the number of items and the average covariance within the test.

Coefficient Alpha - 1

Consider a test made up of k items with an average intercorrelation r.

1) What is the correlation of this test with another test sampled from the same domain of items?
2) What is the correlation of this test with the domain?

	Test 1	Test 2
Test 1	V_1	C_{12}
Test 2	C_{12}	V_2

Let r_1 be the average correlation of items within test 1
Let r_2 be the average correlation of items within test 2
Let r_{12} be the average intercorrelation of items between the two tests.

$$r_{x_1x_2} = \frac{C_{12}}{\sqrt{V_1 * V_2}}$$

	Test 1	Test 2
Test 1	$V_1 = k * [1+(k-1) * r_1]$	$C_{12} = k*k* r_{12}$
Test 2	$C_{12} = k*k* r_{12}$	$V_2 = k * [1+(k-1) * r_2]$

$$r_{x_1x_2} = \frac{k*k*\square\, r_{12}}{\sqrt{k\square * \square [1+(k-1)\square * r_1]\square * k\square * \square [1+(k-1)\square * r_2]\square}}$$

But, since the two tests are composed of randomly equivalent items, $r_1 = r_2 = r$ and

$$r_{x_1x_2} = \frac{k*\square\, r}{1+(k-1)r} = \text{alpha} = \square$$

Note that is the same as the squared correlation of a test with the domain. Alpha is the correlation of a test with a test just like it, and is the percentage of test variance which is domain variance.

Internal Consistency and Coefficient alpha - 2

Consider a test made up of k items with average variance v_i. What is the correlation of this test with another test sampled from the same domain of items?

	Test 1	Test 2
Test 1	V_1	C_{12}
Test 2	C_{12}	V_2

What is the correlation of this test with the domain?

Let V_t be the total test variance for Test 1= $V_1 = V_2$
Let v_i be the average variance of an item within the test.

$$r_{x_1 x_2} = \frac{C_{12}}{\sqrt{V_1 * V_2}}$$

We need to estimate the covariance with the other test:

	Test 1	Test 2
Test 1	$V_1 = k * [v_i + (k-1) * c_1]$	$C_{12} = k * k * r_{12}$
Test 2	$C_{12} = k^2\ c_{12},$	$V_2 = k * [v_i + (k-1) * c_2]$

$C_{12} = k^2\ c_{12}$, but what is the average c_{12}?

$$V_t = V_1 = V_2 \implies c_1 = c_2 = c_{12} \implies$$

$$c_1 = \frac{V_t - \sum v_i}{k*(k-1)} = \text{average covariance}$$

$$C_{12} = k^2\ c_{12} \implies C_{12} = k^2 * \frac{V_t - \sum v_i}{k*(k-1)}$$

$$r_{x_1 x_2} = \frac{k^2 * \frac{V_t - \sum v_i}{k*(k-1)}}{V_t} = \frac{V_t - \sum v_i}{V_t} * \frac{k}{k-1}$$

This allows us to find coefficient alpha without finding the average interitem correlation.

The effect of test length of internal consistency reliability.

	average r	Average r
Number of Items	0.2	0.1
1	0.20	0.10
2	0.33	0.18
4	0.50	0.31
8	0.67	0.47
16	0.80	0.64
32	0.89	0.78
64	0.94	0.88
128	0.97	0.93

Estimates of reliability reflect both the length of the test as well as the average inter-item correlation. To report the internal consistency of a domain (rather than a specific test with a specific length, it is possible to report the "$alpha_1$" for the test.

$$\text{Average interitem } r = alpha_1 = \frac{alpha}{alpha + k*(1 - alpha)}$$

This allows us to find the average internal consistency of a scale independent of test length.

because $\alpha = \frac{V_t - \sum v_i}{V_t} * \frac{k}{k-1}$ is easy to estimate from the basic test statistics and is an estimate of the amount of test variance that is construct related, it should be reported whenever a particular inventory is used.

Coefficients Alpha, Beta and Omega - 1

Components of variance associated with a test score include general test variance, group variance, specific item variance, and error variance.

General	Group	Specific	Error
Reliable Variance			
Common Shared Variance			

Coefficient alpha is the average of all possible splits, and over estimates the general and underestimates the total common variance. It is a lower bound estimate of reliable variance.

Now, consider a test with general and group variance. Each Subtest has general variance but also has Group, Specific, and Error. The subtests share only general variance. How do we estimate the amount of General variance? What would be to correlation of this test with another test with the same general structure, but with different group structures? Find the two most unrelated subtests within each test.

	Subtest A-1	Subtest A-2	Subtest B-3	Subtest B-4
Subtest A-1	$g+G_1+S+E$	g	g	g
Subtest A-2	g	$g+G_2+S+E$	g	g
Subtest B-3	g	g	$g+G_3+S+E$	g
Subtest B-4	g	g		$g+G_4+S+E$

$$r_{ab} = \frac{C_{ab}}{\sqrt{V_a * V_b}} = \frac{4g}{\sqrt{2*(g+G_i+S+E+g)*2*(g+G_i+S+E+g)}}$$

$$\frac{2g}{g+G_1+S+E+g} = \frac{2r_{a1a2}}{1+r_{a1a2}} = \text{“Coefficient Beta”}$$

Coefficient beta is the worst split half reliability and is thus an estimate of the general saturation of the test.

Coefficients Alpha, Beta and Omega - 2

Consider a test with two subtests which are maximally different (the worst split half). What is the predicted correlation with another test formed in the same way?

	Subtest A-1	Subtest A-2	Subtest B-3	Subtest B-4
Subtest A-1	$g+G_1+S+E$	g	g	g
Subtest A-2	g	$g+G_2+S+E$	g	g
Subtest B-3	g	g	$g+G_3+S+E$	g
Subtest B-4	g	g		$g+G_4+S+E$

		Test Size = 10 items		Test Size = 20 items	
General Factor	Group Factor	Alpha	Beta	Alpha	Beta
0.25	0.00	0.77	0.77	0.87	0.87
0.20	0.05	0.75	0.71	0.86	0.83
0.15	0.10	0.73	0.64	0.84	0.78
0.10	0.15	0.70	0.53	0.82	0.69
0.05	0.20	0.67	0.34	0.80	0.51
0.00	0.25	0.63	0.00	0.77	0.00

Note that although alpha is relatively insensitive to the relative contributions of group and general factor, beta is very sensitive. Alpha, however, can be found from item and test statistics, beta needs to be estimated by finding the worst split half. Such an estimate is computationally much more difficult.

Omega, a more general estimate, based upon the factor structure of the test, allows for bette estimate of the first factor saturation.

Generalizabilty Theory
Reliability across facets:

The consistency of Individual Differences across facets may be assessed by analysing variance components associated with each facet. i.e., what amount of variance is associated with a particular facet across which one wants to generalize?

Facets of reliability

Across Items	Domain Sampling Internal Consistency
Across Time	Temporal Stability
Across Forms	Alternate Form Reliability
Across Raters	Inter-rater agreement
Across Situations	Situational Stability
Across "Tests" (facets unspecified)	Parallel Test reliability

Generalizability theory is a decomposition of variance components to estimate sources of variance across which one wants to generalize.

All of these conventional approaches are concerned with generalizing about <u>individual differences</u> (in response to an item, time, form, rater, or situation) between people. Thus, the emphasis is upon consistency of rank orders. Classical reliability is a function of large between subject variability and small within subject variability. It is unable to estimate the within subject precision.

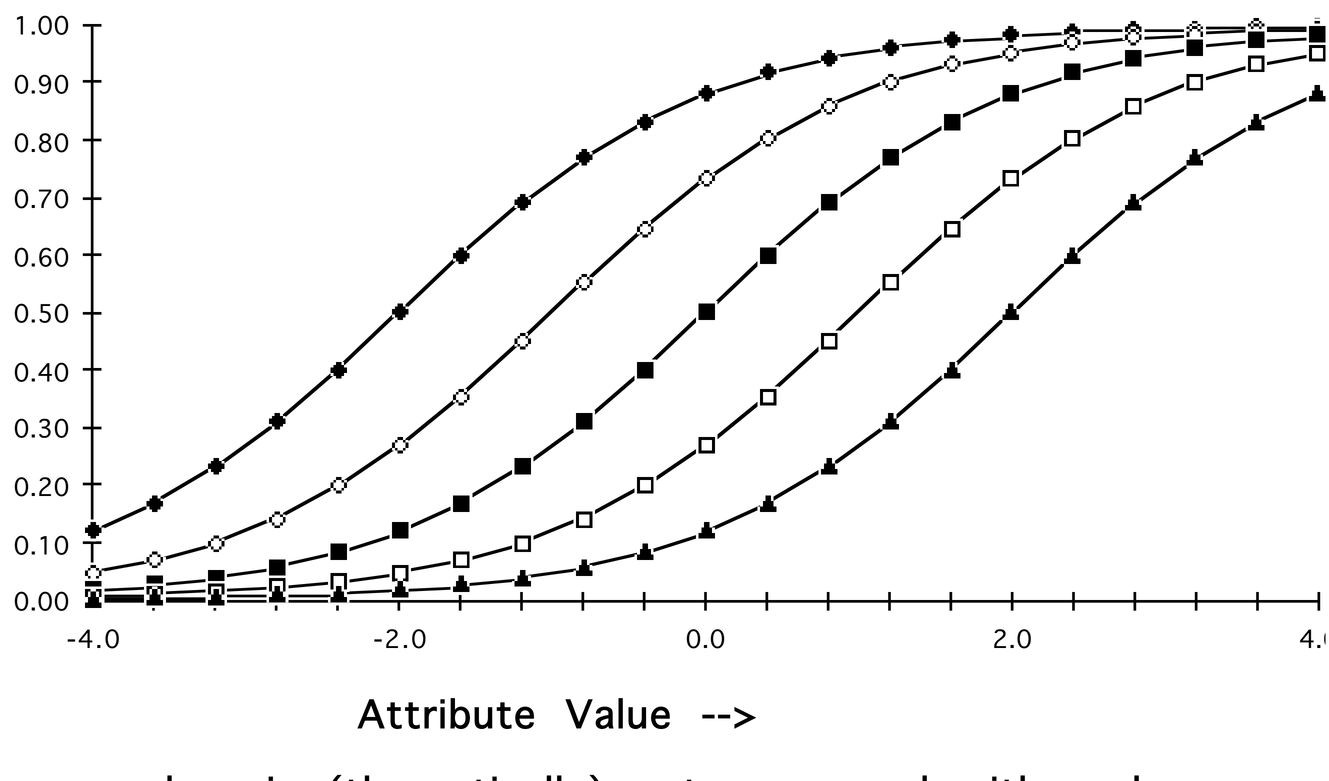
An alternative method (Latent Response Theory or Item Response Theory) is to determine the <u>precision</u> of the estimate of a particular person's position on a latent variable.

Item Response Theory - 1

A model for item response as a function of increasing level of subject ability and increasing levels of item difficulty. This model estimates the probability of making a particular response (generally, correct or incorrect) as a joint function of the subject's value on a latent attribute dimension, and the difficulty (item endorsement rate) of a particular item.

Model 1: the Rasch model: Probability of endorsing an item given ability (ø) and difficulty (diff) :

$$P(y|ø,\text{diff}) = \frac{1}{1+e^{(\text{diff}-ø)}}$$



This procedure is (theoretically) not concerned with rank orders of respondents, but rather with the error of estimate for a particular respondent. This technique allows for computerized adaptive testing.

.

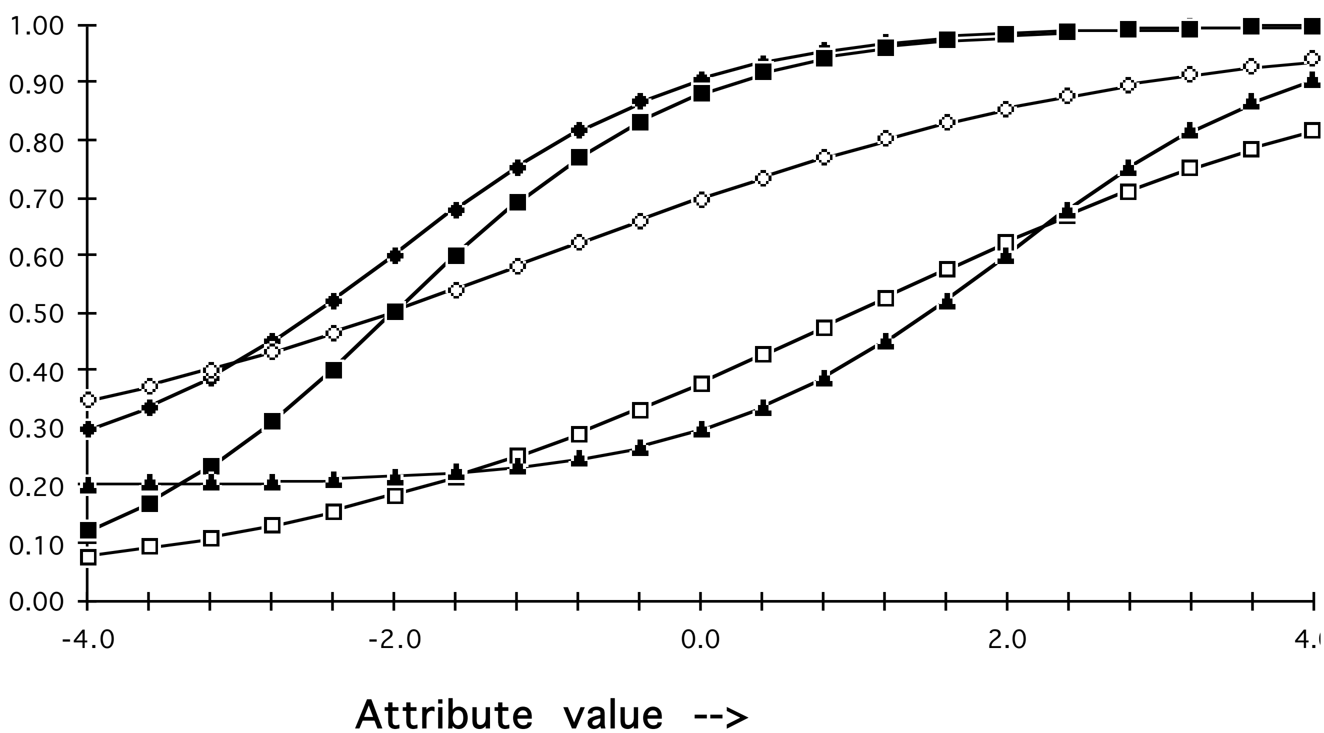
Item Response Theory - 2

A model for item response as a function of increasing level of subject ability and increasing levels of item difficulty.

Model 2: the 3 parameter model: Probability of endorsing an item given ability (ø) ,difficulty (diff), guessing (guessing), and item discrimination sensitivity:

$$P(y|\o,\mathrm{diff},\mathrm{guess},\mathrm{sensitivity}) =$$

$$\mathrm{guessing}+\frac{\mathrm{guessing}}{1+e^{\mathrm{sensitivity}\,*(\mathrm{diff}-\o)}}$$



Note that with this model, even though the probability of item endorsement for a particular item may be a monotonic function of attribute value, item endorsement probabilities for different items may be a non-monotonic function of the attribute.