Item Response Theory

PLAD 8500: Measurement

Item response theory (IRT)

The name “IRT” is a bit unfortunate because it is a measurement model and not in and of itself a theory.

Item response theory (IRT)

The name "IRT" is a bit unfortunate because it is a measurement model and not in and of itself a theory.

We usually think of IRT this way:

Takes binary items only	$\rightarrow$	Spits out an continuous latent variable

Item response theory (IRT)

The name "IRT" is a bit unfortunate because it is a measurement model and not in and of itself a theory.

We usually think of IRT this way:

Takes		Spits out an
binary	$\rightarrow$	continuous
items only		latent variable

But there are versions of IRT that

Item response theory (IRT)

The name “IRT” is a bit unfortunate because it is a measurement model and not in and of itself a theory.

We usually think of IRT this way:

Takes binary items only	$\rightarrow$	Spits out an continuous latent variable

But there are versions of IRT that

- Take ordinal or nominal items (only)

Item response theory (IRT)

The name "IRT" is a bit unfortunate because it is a measurement model and not in and of itself a theory.

We usually think of IRT this way:

Takes binary items only	$\rightarrow$	Spits out an continuous latent variable

But there are versions of IRT that

- Take ordinal or nominal items (only)
- Take ordinal and nominal and binary items together

Item response theory (IRT)

The name "IRT" is a bit unfortunate because it is a measurement model and not in and of itself a theory.

We usually think of IRT this way:

Takes binary items only	$\rightarrow$	Spits out an continuous latent variable

But there are versions of IRT that

- Take ordinal or nominal items (only)
- Take ordinal and nominal and binary items together
- Also take continuous, count, proportion, etc. all together

Item response theory (IRT)

The name "IRT" is a bit unfortunate because it is a measurement model and not in and of itself a theory.

We usually think of IRT this way:

Takes binary items only	$\rightarrow$	Spits out an continuous latent variable

But there are versions of IRT that

- Take ordinal or nominal items (only)
- Take ordinal and nominal and binary items together
- Also take continuous, count, proportion, etc. all together
- Embed a structural equation model

Item response theory (IRT)

The name "IRT" is a bit unfortunate because it is a measurement model and not in and of itself a theory.

We usually think of IRT this way:

Takes binary items only	$\rightarrow$	Spits out an continuous latent variable

But there are versions of IRT that

- Take ordinal or nominal items (only)
- Take ordinal and nominal and binary items together
- Also take continuous, count, proportion, etc. all together
- Embed a structural equation model
- Use **time series** data

Canonical Application of IRT: Grading a test

$X_1, X_2, \ldots, X_k$ are questions (items) on a test.

Canonical Application of IRT: Grading a test

$X_1, X_2, \ldots, X_k$ are questions (items) on a test.

The items are binary:

- $X_{ij} = 1$ if student i gets question j correct,
- otherwise $X_{ij} = 0$.

Canonical Application of IRT: Grading a test

$X_1, X_2, \ldots, X_k$ are questions (items) on a test.

The items are binary:

- $X_{ij} = 1$ if student i gets question j correct,
- otherwise $X_{ij} = 0$.

θ_i is student i's latent ability.

Canonical Application of IRT: Grading a test

$X_1, X_2, \ldots, X_k$ are questions (items) on a test.

The items are binary:

- $X_{ij} = 1$ if student i gets question j correct,
- otherwise $X_{ij} = 0$.

θ_i is student i's latent ability.

To **measure ability**, why not simply take a sum of the Xs?

Canonical Application of IRT: Grading a test

$X_1, X_2, \ldots, X_k$ are questions (items) on a test.

The items are binary:

- $X_{ij} = 1$ if student i gets question j correct,
- otherwise $X_{ij} = 0$.

θ_i is student i's latent ability.

To **measure ability**, why not simply take a sum of the Xs?
Because some Xs are more informative about θ than others.

Canonical Application of IRT: Grading a test

$X_1, X_2, \ldots, X_k$ are questions (items) on a test.

The items are binary:

- $X_{ij} = 1$ if student i gets question j correct,
- otherwise $X_{ij} = 0$.

θ_i is student i's latent ability.

To **measure ability**, why not simply take a sum of the Xs?
Because some Xs are more informative about θ than others.

IRT weights the items on two criteria:

1. The **difficulty** of each question,

Canonical Application of IRT: Grading a test

$X_1, X_2, \ldots, X_k$ are questions (items) on a test.

The items are binary:

- $X_{ij} = 1$ if student i gets question j correct,
- otherwise $X_{ij} = 0$.

θ_i is student i's latent ability.

To **measure ability**, why not simply take a sum of the Xs?
Because some Xs are more informative about θ than others.

IRT weights the items on two criteria:

1. The **difficulty** of each question,
2. and the ability of a question to **discriminate** between high and low ability students.

Three topics we need to review

I love item response theory (IRT). I think it provides a great balance between theoretical modeling and flexible measurement.

Three topics we need to review

I love item response theory (IRT). I think it provides a great balance between theoretical modeling and flexible measurement.

IRT comes from psychometrics, and it is *underused* because psychologists seem unwilling to apply it more broadly than the canonical problems, and economists have little use for it.

Three topics we need to review

I love item response theory (IRT). I think it provides a great balance between theoretical modeling and flexible measurement.

IRT comes from psychometrics, and it is *underused* because psychologists seem unwilling to apply it more broadly than the canonical problems, and economists have little use for it.

But before we can delve into this topic we must review three topics:

- Bayes' rule and proportionality
- Confirmatory factor analysis and path diagrams
- Generalized linear models (GLM)

Review: Bayes' rule

X — a binary random variable.
Does the student get the question right or wrong?

Review: Bayes' rule

X — a binary random variable.
Does the student get the question right or wrong?

θ — a continuous random variable that influences X.
The student's ability.

Review: Bayes' rule

X — a binary random variable.
Does the student get the question right or wrong?

θ — a continuous random variable that influences X.
The student's ability.

$P(X|\theta)$ – the conditional probability of X given θ
What is the probability that a student of a particular ability level gets the question right?

Review: Bayes' rule

X — a binary random variable.
Does the student get the question right or wrong?

θ — a continuous random variable that influences X.
The student's ability.

$P(X|\theta)$ – the conditional probability of X given θ
What is the probability that a student of a particular ability level gets the question right?

$P(\theta|X)$ – the conditional probability of θ given X
What is the probability that a student who gets the question right has a particular ability level?

Review: Bayes' rule

X — a binary random variable.
Does the student get the question right or wrong?

θ — a continuous random variable that influences X.
The student's ability.

$P(X|\theta)$ – the conditional probability of X given θ
What is the probability that a student of a particular ability level gets the question right?

$P(\theta|X)$ – the conditional probability of θ given X
What is the probability that a student who gets the question right has a particular ability level?

Our goal is to estimate these two conditional probabilites.

Review: Bayes' rule

Suppose that we know $P(X|\theta)$. We can find the other conditional probability by using Bayes' rule:

Review: Bayes' rule

Suppose that we know $P(X|\theta)$. We can find the other conditional probability by using Bayes' rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

Review: Bayes' rule

Suppose that we know $P(X|\theta)$. We can find the other conditional probability by using Bayes' rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

In general, we don't know $P(X)$, but we know $P(X|\theta)$ for any value of θ.

Review: Bayes' rule

Suppose that we know $P(X|\theta)$. We can find the other conditional probability by using Bayes' rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

In general, we don't know $P(X)$, but we know $P(X|\theta)$ for any value of θ.

Suppose there were two values of θ:

low ability θ_L and high ability θ_H.

Review: Bayes' rule

Suppose that we know $P(X|\theta)$. We can find the other conditional probability by using Bayes' rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}.$$

In general, we don't know $P(X)$, but we know $P(X|\theta)$ for any value of θ.

Suppose there were two values of θ:

low ability θ_L and high ability θ_H.

Then we could rewrite the denominator as

$$P(X) = P(X|\theta_L)P(\theta_L) + P(X|\theta_H)P(\theta_H).$$

Review: Bayes' rule

Suppose there were ten values:

$$\theta_1, \ldots, \theta_{10}$$

Then we could rewrite the denominator as

$$P(X) = \sum_{i=1}^{10} P(X|\theta_i)P(\theta_i).$$

Review: Bayes' rule

Suppose there were ten values:

$$\theta_1, \dots, \theta_{10}$$

Then we could rewrite the denominator as

$$P(X) = \sum_{i=1}^{10} P(X|\theta_i)P(\theta_i).$$

But if θ is continuous, there are infinitely many values. The infinite analogue of a sum is an integral. So in this case:

$$P(X) = \int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta.$$

Review: Bayes' rule

So we can rewrite Bayes' rule as

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta}.$$

Review: Bayes' rule

So we can rewrite Bayes' rule as

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta}.$$

IRT (and most Bayesian methods) use a dirty trick that depends on **proportionality**. Two functions are proportional if one is a multiple of the other.

Review: Bayes' rule

So we can rewrite Bayes' rule as

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta}.$$

IRT (and most Bayesian methods) use a dirty trick that depends on **proportionality**. Two functions are proportional if one is a multiple of the other.

Example:

$$f(x) = x^2 + 1, \quad g(x) = 3x^2 + 3, \quad h(x) = -.75x^2 - .75$$

are all proportional because

$$g(x) = 3f(x), \qquad h(x) = -.75f(x).$$

Review: Bayes' rule

So we can rewrite Bayes' rule as

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta}.$$

IRT (and most Bayesian methods) use a dirty trick that depends on **proportionality**. Two functions are proportional if one is a multiple of the other.

<u>Example</u>:

$$f(x) = x^2 + 1, \quad g(x) = 3x^2 + 3, \quad h(x) = -.75x^2 - .75$$

are all proportional because

$$g(x) = 3f(x), \qquad h(x) = -.75f(x).$$

The sign $\propto$ means "proportional to."

$$f(x) \propto g(x), \qquad f(x) \propto h(x).$$

Review: Bayes' rule

It turns out that

$$P(X) = \int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta$$

is just equal to a single, scalar value. *We don't need to know what this value is*. Since it is scalar, we can rewrite Bayes' rule again like this:

$$P(\theta|X) \propto P(X|\theta)P(\theta).$$

Review: Bayes' rule

It turns out that

$$P(X) = \int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta$$

is just equal to a single, scalar value. *We don't need to know what this value is.* Since it is scalar, we can rewrite Bayes' rule again like this:

$$P(\theta|X) \propto P(X|\theta)P(\theta).$$

This setup gives us a curve for $P(\theta|X)$ that has the right shape, but the **wrong scale**.

Review: Bayes' rule

It turns out that

$$P(X) = \int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta$$

is just equal to a single, scalar value. *We don't need to know what this value is.* Since it is scalar, we can rewrite Bayes' rule again like this:

$$P(\theta|X) \propto P(X|\theta)P(\theta).$$

This setup gives us a curve for $P(\theta|X)$ that has the right shape, but the **wrong scale**.

The dirty trick we use is **drawing θ values from this curve**.

Review: Bayes' rule

It turns out that

$$P(X) = \int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta$$

is just equal to a single, scalar value. *We don't need to know what this value is*. Since it is scalar, we can rewrite Bayes' rule again like this:

$$P(\theta|X) \propto P(X|\theta)P(\theta).$$

This setup gives us a curve for $P(\theta|X)$ that has the right shape, but the **wrong scale**.

The dirty trick we use is **drawing θ values from this curve**. We know that this technique

- does not change the maximum or mean,

Review: Bayes' rule

It turns out that

$$P(X) = \int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta$$

is just equal to a single, scalar value. *We don't need to know what this value is.* Since it is scalar, we can rewrite Bayes' rule again like this:

$$P(\theta|X) \propto P(X|\theta)P(\theta).$$

This setup gives us a curve for $P(\theta|X)$ that has the right shape, but the **wrong scale**.

The dirty trick we use is **drawing θ values from this curve**. We know that this technique

- does not change the maximum or mean,
- and the 2.5% and 97.5% percentiles of simulated θ values are a correct estimate of the 95% "credible" (like a confidence) interval.

Review: Bayes' rule

So what? What does all this technical Bayes' stuff mean?

Review: Bayes' rule

So what? What does all this technical Bayes' stuff mean?

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

It means that if we can figure out **two things**:

Review: Bayes' rule

So what? What does all this technical Bayes' stuff mean?

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

It means that if we can figure out **two things**:

1. $P(X|\theta)$, the probability of a correct answer given an ability level,

Review: Bayes' rule

So what? What does all this technical Bayes' stuff mean?

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

It means that if we can figure out **two things**:

1. $P(X|\theta)$, the probability of a correct answer given an ability level,
2. and $P(\theta)$, a reasonable prior expectation for ability,

Review: Bayes' rule

So what? What does all this technical Bayes' stuff mean?

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

It means that if we can figure out **two things**:

1. $P(X|\theta)$, the probability of a correct answer given an ability level,
2. and $P(\theta)$, a reasonable prior expectation for ability,

then it's easy to estimate $P(\theta|X)$ by **multiplying these two functions together**.

Review: Bayes' rule

So what? What does all this technical Bayes' stuff mean?

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

It means that if we can figure out **two things**:

1. $P(X|\theta)$, the probability of a correct answer given an ability level,
2. and $P(\theta)$, a reasonable prior expectation for ability,

then it's easy to estimate $P(\theta|X)$ by **multiplying these two functions together**.

Usually, we assume that every student has a prior $P(\theta)$ that is **standard normal** (like assuming every student is average). Then the test answers let us update that belief.

Review: Bayes' rule

So what? What does all this technical Bayes' stuff mean?

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

It means that if we can figure out **two things**:

1. $P(X|\theta)$, the probability of a correct answer given an ability level,
2. and $P(\theta)$, a reasonable prior expectation for ability,

then it's easy to estimate $P(\theta|X)$ by **multiplying these two functions together**.

Usually, we assume that every student has a prior $P(\theta)$ that is **standard normal** (like assuming every student is average). Then the test answers let us update that belief.

All we need now is a model for $P(X|\theta)$!

Review: CFA and path diagrams

Let X_1, X_2, ..., X_k be **items** (that's the language of IRT – but these are also indicators, measures, observed variables, etc.)

Review: CFA and path diagrams

Let X_1, X_2, …, X_k be **items** (that's the language of IRT – but these are also indicators, measures, observed variables, etc.)

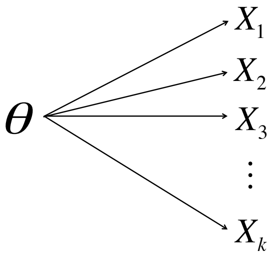
Let θ be the latent variable (the factor, the score, etc.)

Review: CFA and path diagrams

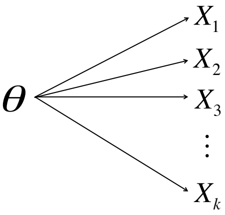
Let X_1, X_2, ..., X_k be **items** (that's the language of IRT – but these are also indicators, measures, observed variables, etc.)

Let θ be the latent variable (the factor, the score, etc.)

Remember, confirmatory factor analysis is built on a path diagram:

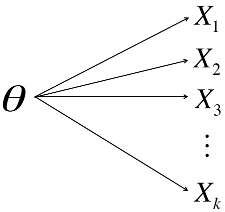


Review: CFA and path diagrams



That means that θ is the independent variable and the items are the dependent variables.

Review: CFA and path diagrams



That means that θ is the independent variable and the items are the dependent variables.

This diagram implies a system of equations as follows:

$$\begin{cases} X_1 = \alpha_1 + \beta_1\theta + \varepsilon_1, \\ X_2 = \alpha_2 + \beta_2\theta + \varepsilon_2, \\ \vdots, \\ X_k = \alpha_k + \beta_k\theta + \varepsilon_k. \end{cases}$$

Review: CFA and path diagrams

$$\begin{cases} X_1 = \alpha_1 + \beta_1\theta + \varepsilon_1, \\ X_2 = \alpha_2 + \beta_2\theta + \varepsilon_2, \\ \vdots, \\ X_k = \alpha_k + \beta_k\theta + \varepsilon_k. \end{cases}$$

Since each X is a dependent variable, and θ is an independent variable, we can write each model as

$$f(X_k|\theta),$$

Review: CFA and path diagrams

$$
\begin{cases}
X_1 = \alpha_1 + \beta_1\theta + \varepsilon_1, \\
X_2 = \alpha_2 + \beta_2\theta + \varepsilon_2, \\
\vdots, \\
X_k = \alpha_k + \beta_k\theta + \varepsilon_k.
\end{cases}
$$

Since each X is a dependent variable, and θ is an independent variable, we can write each model as

$$f(X_k|\theta),$$

which is exactly what we need to solve Bayes' rule for the posterior estimate of θ.

Review: GLM

What is a really common way that social scientists can model

$$f(X_k|\theta),$$

if the outcome X is binary?

Review: GLM

What is a really common way that social scientists can model

$$f(X_k|\theta),$$

if the outcome X is binary?

Logistic (logit) regression.

Review: GLM

What is a really common way that social scientists can model

$$f(X_k|\theta),$$

if the outcome X is binary?

Logistic (logit) regression. Probit regression works too, but most IRT users work with logit, and that will be our focus too.

Review: GLM

What is a really common way that social scientists can model

$$f(X_k|\theta),$$

if the outcome X is binary?

Logistic (logit) regression. Probit regression works too, but most IRT users work with logit, and that will be our focus too.

Logit and probit are examples of generalized linear models (GLMs). A GLM has three parts:

Review: GLM

What is a really common way that social scientists can model

$$f(X_k|\theta),$$

if the outcome X is binary?

Logistic (logit) regression. Probit regression works too, but most IRT users work with logit, and that will be our focus too.

Logit and probit are examples of generalized linear models (GLMs). A GLM has three parts:

1. A family that represents the variation of the outcomes. The family has **parameters** that alter the shape of the distribution.

Review: GLM

What is a really common way that social scientists can model

$$f(X_k|\theta),$$

if the outcome X is binary?

Logistic (logit) regression. Probit regression works too, but most IRT users work with logit, and that will be our focus too.

Logit and probit are examples of generalized linear models (GLMs). A GLM has three parts:

1. A family that represents the variation of the outcomes. The family has **parameters** that alter the shape of the distribution.
2. A linear model that contains the independent variables and coefficients.

Review: GLM

What is a really common way that social scientists can model

$$f(X_k|\theta),$$

if the outcome X is binary?

Logistic (logit) regression. Probit regression works too, but most IRT users work with logit, and that will be our focus too.

Logit and probit are examples of generalized linear models (GLMs). A GLM has three parts:

1. A family that represents the variation of the outcomes. The family has **parameters** that alter the shape of the distribution.
2. A linear model that contains the independent variables and coefficients.
3. A link function that allows you to substitute the linear model for one of the family's parameters.

Review: GLM

Logistic regression:

Review: GLM

Logistic regression:

Family: the Bernoulli distribution

$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$

Review: GLM

Logistic regression:

Family: the Bernoulli distribution

$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$

Parameter: p_i – the probability that observation i is a 1. p_i must be between 0 and 1.

Review: GLM

Logistic regression:

Family: the Bernoulli distribution

$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$

Parameter: p_i – the probability that observation i is a 1. p_i must be between 0 and 1.

Linear model: denoted y_i^* and allowed to take on all real numbers,

$$y_i^* = \alpha + \beta_1 x_{i1} + \ldots \beta_k x_{ik}$$

Review: GLM

Logistic regression:

Family: the Bernoulli distribution

$$f(y_i|p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}.$$

Parameter: p_i – the probability that observation i is a 1. p_i must be between 0 and 1.

Linear model: denoted y_i^* and allowed to take on all real numbers,

$$y_i^* = \alpha + \beta_1 x_{i1} + \ldots \beta_k x_{ik}$$

Link function: a function that converts the domain of y^* (all reals) to the domain of the parameter (between 0 and 1).

Review: GLM

Logistic regression:

Family: the Bernoulli distribution

$$f(y_i|p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}.$$

Parameter: p_i – the probability that observation i is a 1. p_i must be between 0 and 1.

Linear model: denoted y_i^* and allowed to take on all real numbers,

$$y_i^* = \alpha + \beta_1 x_{i1} + \dots \beta_k x_{ik}$$

Link function: a function that converts the domain of y^* (all reals) to the domain of the parameter (between 0 and 1).

For logit, we use the logistic CDF:

$$p_i = \frac{1}{1 + e^{-y_i^*}}.$$

Review: GLM

Suppose that the only independent variable were θ. Then we could write

$$y_i^* = b_0 + b_1\theta.$$

Review: GLM

Suppose that the only independent variable were θ. Then we could write

$$y_i^* = b_0 + b_1\theta.$$

Equivalently we can write this as

$$y_i^* = \alpha(\theta - \beta)$$

where $\alpha = b_1$ and $\beta = -b_1/b_0$.

Review: GLM

Suppose that the only independent variable were θ. Then we could write

$$y_i^* = b_0 + b_1\theta.$$

Equivalently we can write this as

$$y_i^* = \alpha(\theta - \beta)$$

where $\alpha = b_1$ and $\beta = -b_1/b_0$.

Then the probabilities are

$$P(X = 1|\theta) = p_i = \frac{1}{1 + \exp\Big(- \alpha(\theta - \beta)\Big)},$$

$$P(X = 0|\theta) = 1 - p_i = 1 - \frac{1}{1 + \exp\Big(- \alpha(\theta - \beta)\Big)}.$$

Test curves

An **IRT test curve** looks like this:

$$P(X_i = 1|\theta_i) = \frac{1}{1 + \exp\Big(- \alpha(\theta_i - \beta)\Big)}.$$

Test curves

An **IRT test curve** looks like this:

$$P(X_i = 1|\theta_i) = \frac{1}{1 + \exp\Big(-\alpha(\theta_i - \beta)\Big)}.$$

Suppose that we already know θ_i. Then to get ML estimates of α and β, we estimate a logistic regression.

Test curves

An **IRT test curve** looks like this:

$$P(X_i = 1|\theta_i) = \frac{1}{1 + \exp\Big(-\alpha(\theta_i - \beta)\Big)}.$$

Suppose that we already know θ_i. Then to get ML estimates of α and β, we estimate a logistic regression.

The parameters:

Test curves

An **IRT test curve** looks like this:

$$P(X_i = 1|\theta_i) = \frac{1}{1 + \exp\Big(-\alpha(\theta_i - \beta)\Big)}.$$

Suppose that we already know θ_i. Then to get ML estimates of α and β, we estimate a logistic regression.

The parameters:

- β – DIFFICULTY.

Test curves

An **IRT test curve** looks like this:

$$P(X_i = 1|\theta_i) = \frac{1}{1 + \exp\Big(-\alpha(\theta_i - \beta)\Big)}.$$

Suppose that we already know θ_i. Then to get ML estimates of α and β, we estimate a logistic regression.

The parameters:

- β – DIFFICULTY.
  Represents the level of ability necessary to have a 50/50 chance of getting the problem right. Higher values mean the question is more difficult.

Test curves

An **IRT test curve** looks like this:

$$P(X_i = 1|\theta_i) = \frac{1}{1 + \exp\Big(- \alpha(\theta_i - \beta)\Big)}.$$

Suppose that we already know θ_i. Then to get ML estimates of α and β, we estimate a logistic regression.

The parameters:

- β – DIFFICULTY.
  Represents the level of ability necessary to have a 50/50 chance of getting the problem right. Higher values mean the question is more difficult.
- α – DISCRIMINATION

Test curves

An **IRT test curve** looks like this:

$$P(X_i = 1|\theta_i) = \frac{1}{1 + \exp\Big(- \alpha(\theta_i - \beta)\Big)}.$$

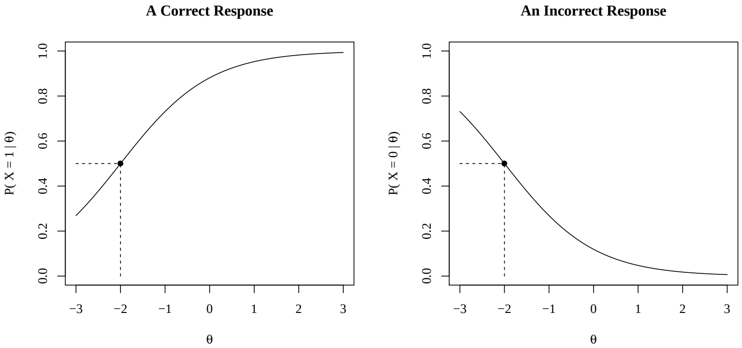
Suppose that we already know θ_i. Then to get ML estimates of α and β, we estimate a logistic regression.

The parameters:

- β – DIFFICULTY.
  Represents the level of ability necessary to have a 50/50 chance of getting the problem right. Higher values mean the question is more difficult.
- α – DISCRIMINATION
  Represents how quickly probabilities go to 0 to the left of the .5 point, and how quickly probabilities go to 1 to the right.

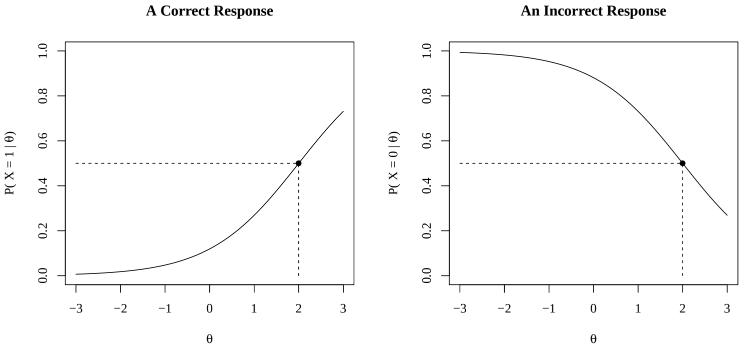
Test Curve: the probability of an item equalling 1 conditional on θ and on the item parameters.

An easy item, $\beta = -2$



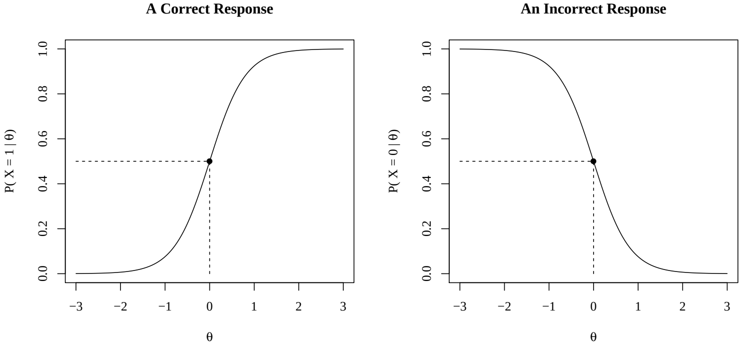
Test Curve: the probability of an item equalling 1 conditional on θ and on the item parameters.

A difficult item, $\beta = 2$



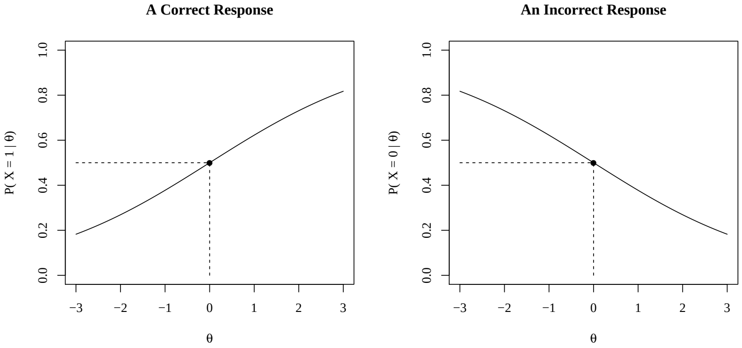
Test Curve: the probability of an item equalling 1 conditional on θ and on the item parameters.

An item that discriminates between high and low ability students well, $\alpha = 1$



Test Curve: the probability of an item equalling 1 conditional on θ and on the item parameters.

An item that discriminates between high and low ability students poorly, $\alpha = 0.1$



Some notes about test curves

Discrimination parameters are like **factor loadings**:

Some notes about test curves

Discrimination parameters are like **factor loadings**:

- The steeper the curve,
- the more the variation of X is explained by θ,
- the better the item fits as a measure of the latent variable.

Some notes about test curves

Discrimination parameters are like **factor loadings**:

- The steeper the curve,
- the more the variation of X is explained by θ,
- the better the item fits as a measure of the latent variable.

Low discrimination = a bad item. Like including a Shakespeare question on a math test. *Bad students still might get it right, good students still might get it wrong*.

Some notes about test curves

Discrimination parameters are like **factor loadings**:

- The steeper the curve,
- the more the variation of X is explained by θ,
- the better the item fits as a measure of the latent variable.

Low discrimination = a bad item. Like including a Shakespeare question on a math test. *Bad students still might get it right, good students still might get it wrong*.

There are other versions with more/fewer parameters, different distributions. But this setup is most common.

Some notes about test curves

Discrimination parameters are like **factor loadings**:

- The steeper the curve,
- the more the variation of X is explained by θ,
- the better the item fits as a measure of the latent variable.

Low discrimination = a bad item. Like including a Shakespeare question on a math test. *Bad students still might get it right, good students still might get it wrong*.

There are other versions with more/fewer parameters, different distributions. But this setup is most common.

Can be estimated through iterated ML or MCMC.

How to estimate θ?

Start with a prior for the value of θ for every observation i.

How to estimate θ?

Start with a prior for the value of θ for every observation i.

A common approach: everyone's θ has a standard normal prior, independent of one another.

How to estimate θ?

Start with a prior for the value of θ for every observation i.

A common approach: everyone's θ has a standard normal prior, independent of one another.

Then consider item 1 (if we observe $X_{i1} = 1$). **Suppose we know** α_1 **and** β_1. Then we know

$$P(X_{i1} = 1|\theta_i) = \frac{1}{1 + \exp\Big(- \alpha_1(\theta_i - \beta_1)\Big)},$$

How to estimate θ?

Start with a prior for the value of θ for every observation i.

A common approach: everyone's θ has a standard normal prior, independent of one another.

Then consider item 1 (if we observe $X_{i1} = 1$). **Suppose we know** α_1 **and** β_1. Then we know

$$P(X_{i1} = 1|\theta_i) = \frac{1}{1 + \exp\Big(-\alpha_1(\theta_i - \beta_1)\Big)},$$

Update the estimate of θ_i using Bayes' rule:

$$P(\theta_i|X_{i1} = 1) \propto P(X_{i1} = 1|\theta_i)P(\theta_i)$$

How to estimate θ?

Replace the prior $P(\theta_i)$ with the latest posterior: $P(\theta_i|X_{i1})$.

How to estimate θ?

Replace the prior $P(\theta_i)$ with the latest posterior: $P(\theta_i|X_{i1})$.

Then consider item 2 (if we observe $X_{i2} = 0$). **Suppose we know α_2 and β_2**, so we know

$$P(X_{i2} = 0|\theta_i) = 1 - \frac{1}{1 + \exp\Big(- \alpha_2(\theta_i - \beta_2)\Big)}, .$$

How to estimate θ?

Replace the prior $P(\theta_i)$ with the latest posterior: $P(\theta_i|X_{i1})$.

Then consider item 2 (if we observe $X_{i2} = 0$). **Suppose we know α_2 and β_2**, so we know

$$P(X_{i2} = 0|\theta_i) = 1 - \frac{1}{1 + \exp\Big(-\alpha_2(\theta_i - \beta_2)\Big)},.$$

Update the estimate of θ using Bayes' rule:

$$P(\theta_i|X_{i2} = 0) \propto P(X_{i2} = 1|\theta_i)P(\theta_i|X_{i1})$$

How to estimate θ?

Replace the prior $P(\theta_i)$ with the latest posterior: $P(\theta_i|X_{i1})$.

Then consider item 2 (if we observe $X_{i2} = 0$). **Suppose we know α_2 and β_2**, so we know

$$P(X_{i2} = 0|\theta_i) = 1 - \frac{1}{1 + \exp\Big(-\alpha_2(\theta_i - \beta_2)\Big)}, .$$

Update the estimate of θ using Bayes' rule:

$$P(\theta_i|X_{i2} = 0) \propto P(X_{i2} = 1|\theta_i)P(\theta_i|X_{i1})$$

Replace the prior with the latest posterior: $P(\theta_i|X_{i1}, X_{i2})$. Repeat for every item.

How to estimate θ?

The estimate of θ for observation i turns out to be the PRODUCT of

- the (original) prior distribution of θ_i,
- and **every test curve** for observation i.

Just multiply everything together!

How to estimate θ?

The estimate of θ for observation i turns out to be the PRODUCT of

- the (original) prior distribution of θ_i,
- and **every test curve** for observation i.

Just multiply everything together!

The resulting curve has the right shape, wrong scale.

How to estimate θ?

The estimate of θ for observation i turns out to be the PRODUCT of

- the (original) prior distribution of θ_i,
- and **every test curve** for observation i.

Just multiply everything together!

The resulting curve has the right shape, wrong scale.

Use (MCMC or EM) **simulation** to

How to estimate θ?

The estimate of θ for observation i turns out to be the PRODUCT of

- the (original) prior distribution of θ_i,
- and **every test curve** for observation i.

Just multiply everything together!

The resulting curve has the right shape, wrong scale.

Use (MCMC or EM) **simulation** to

- Obtain standard errors, confidence intervals,

How to estimate θ?

The estimate of θ for observation i turns out to be the PRODUCT of

- the (original) prior distribution of θ_i,
- and **every test curve** for observation i.

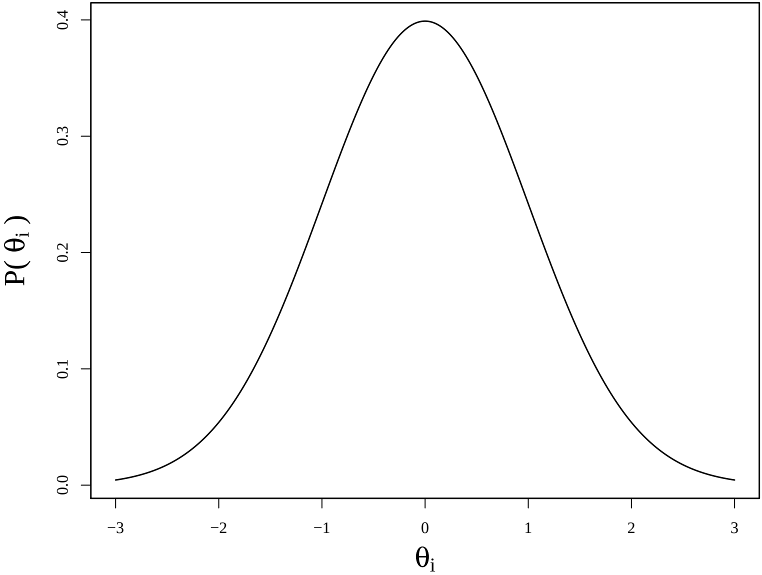
Just multiply everything together!

The resulting curve has the right shape, wrong scale.

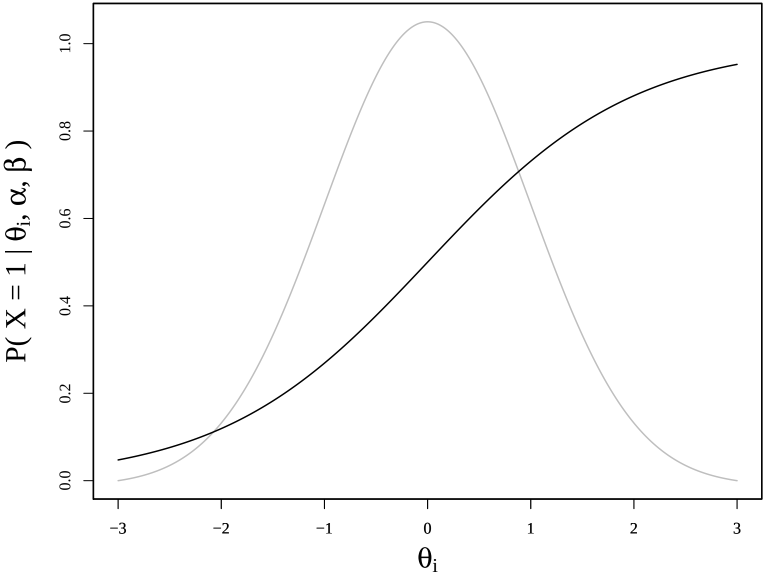
Use (MCMC or EM) **simulation** to

- Obtain standard errors, confidence intervals,
- Go back and forth between estimating test curves and θ until these quantities converge to one answer.

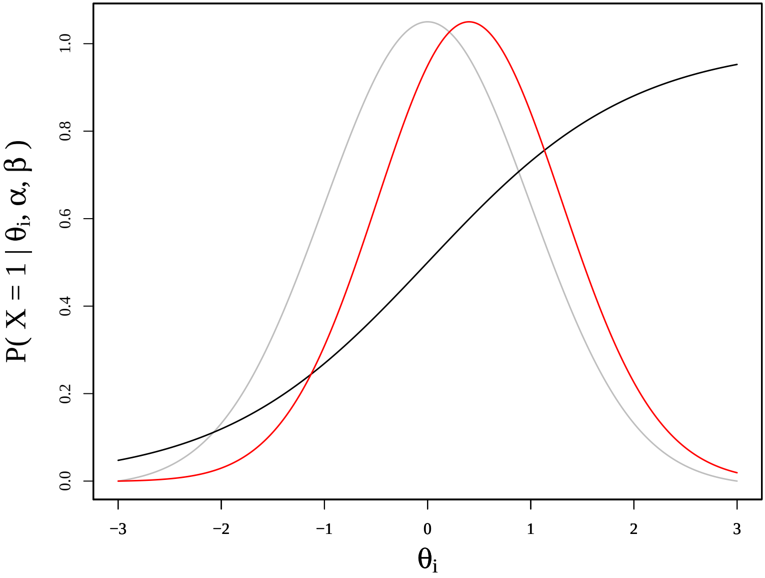
Prior distribution of the latent variable θ_i:



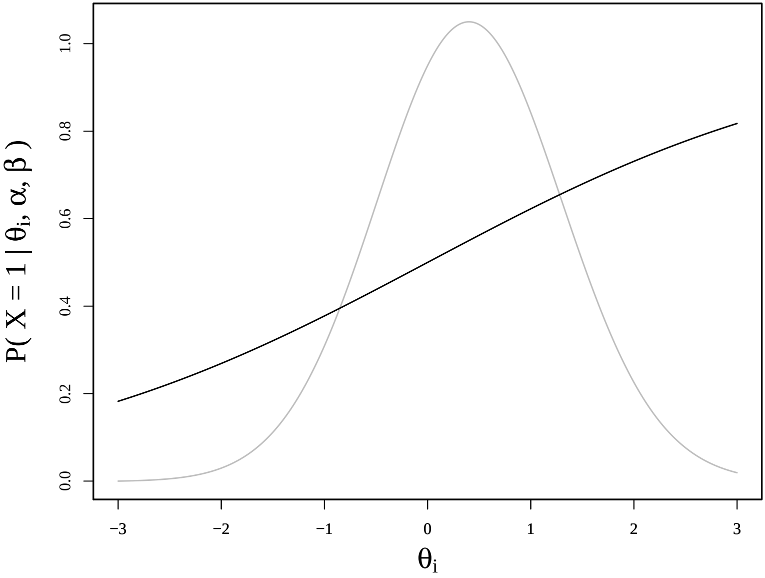
Item 1: medium difficulty, medium discrimination, CORRECT



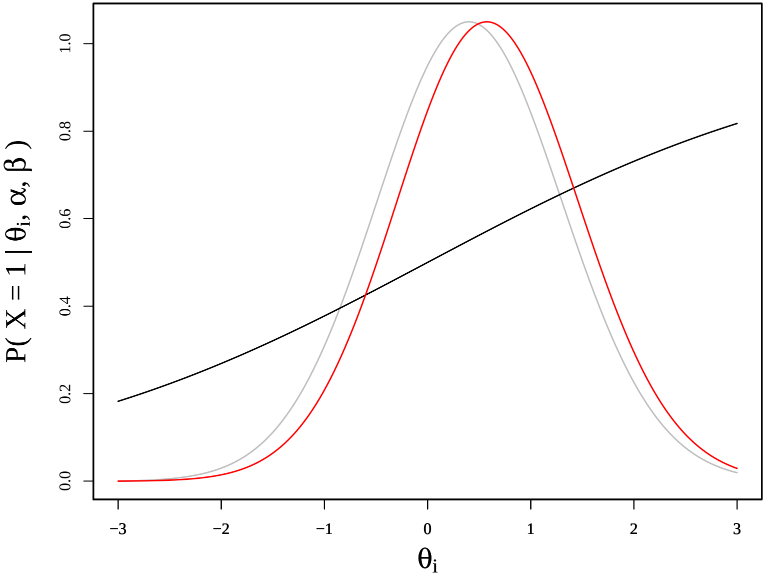
Item 1: medium difficulty, medium discrimination, CORRECT



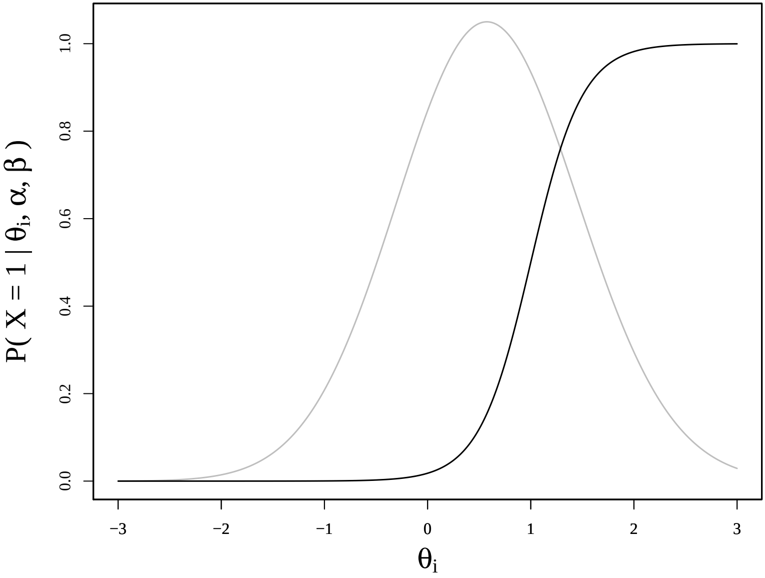
Item 2: medium difficulty, low discrimination, CORRECT



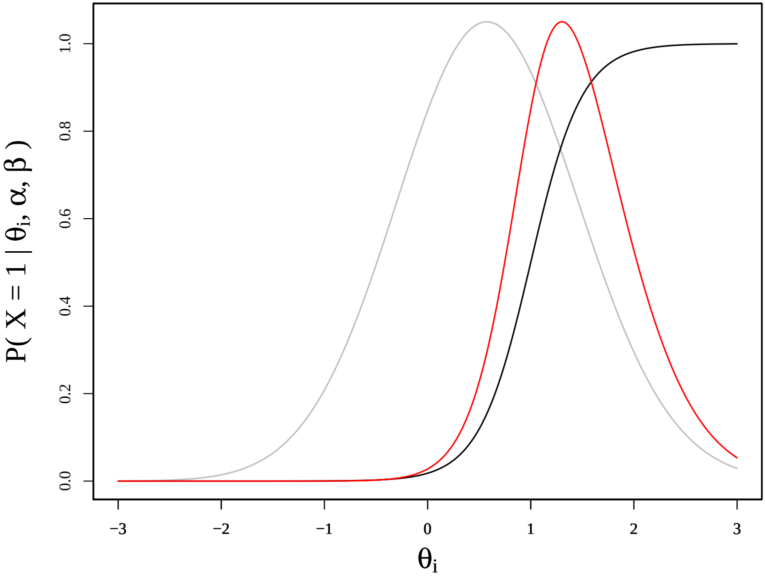
Item 2: medium difficulty, low discrimination, CORRECT



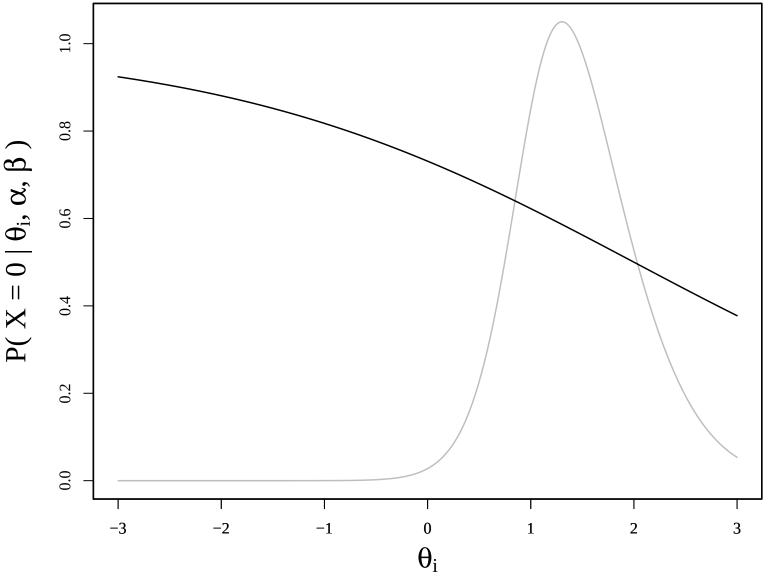
Item 3: high difficulty, high discrimination, CORRECT



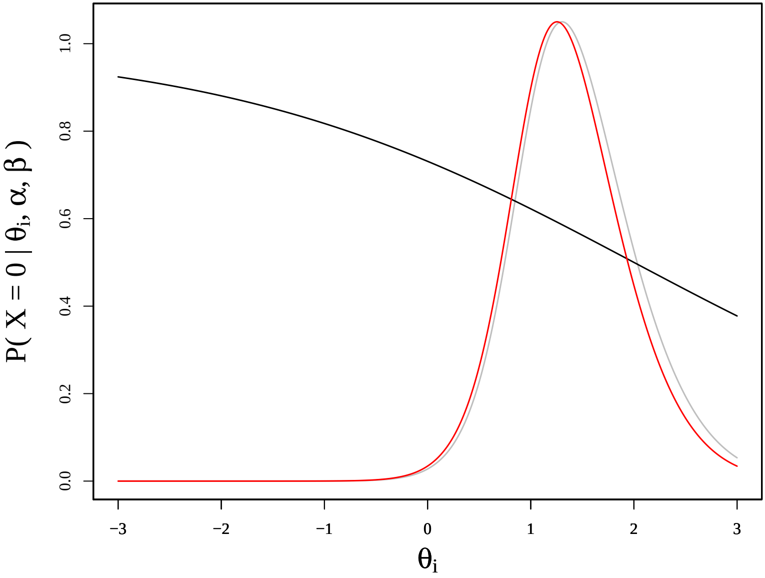
Item 3: high difficulty, high discrimination, CORRECT



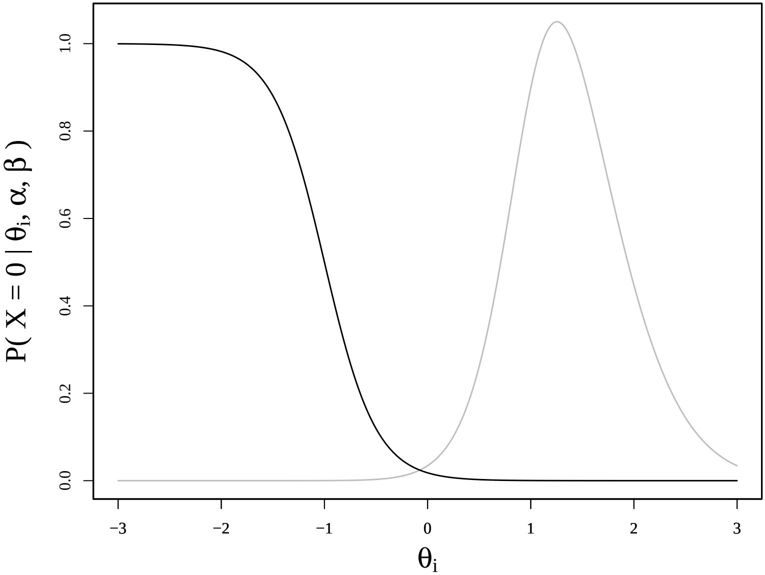
Item 4: high difficulty, low discrimination, INCORRECT



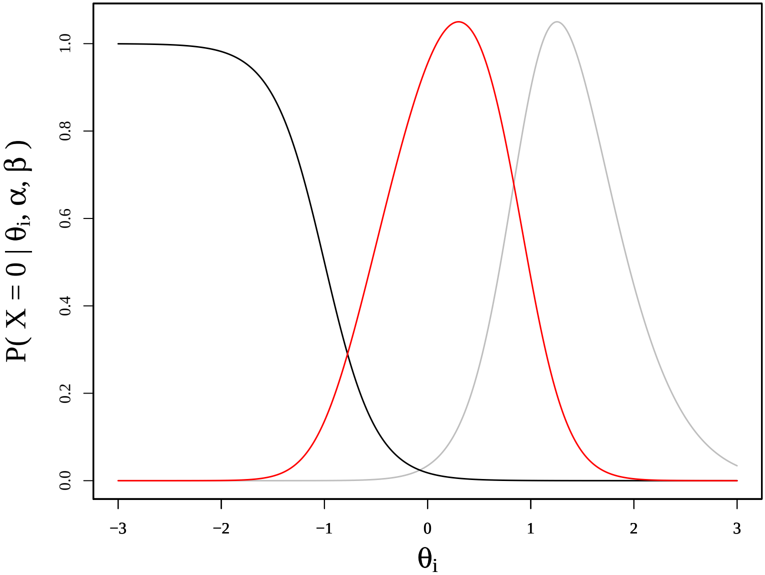
Item 4: high difficulty, low discrimination, INCORRECT



Item 5: low difficulty, high discrimination, INCORRECT



Item 5: low difficulty, high discrimination, INCORRECT



Assumptions

IRT makes a few strong assumptions:

Assumptions

IRT makes a few strong assumptions:

First, observations are independent conditional on θ.
Two students' test answers are related only in so far that the students have similar ability.

Assumptions

IRT makes a few strong assumptions:

First, observations are independent conditional on θ.
Two students' test answers are related only in so far that the students have similar ability.

Second, items are independent conditional on θ.
Getting one question right should not affect the probability of getting another question right, outside of the ability demonstrated in answering both questions.

Assumptions

IRT makes a few strong assumptions:

First, observations are independent conditional on θ.
Two students' test answers are related only in so far that the students have similar ability.

Second, items are independent conditional on θ.
Getting one question right should not affect the probability of getting another question right, outside of the ability demonstrated in answering both questions.

Usually (but not necessarily), the priors $P(\theta_i)$ are assumed to be independent across observations. That implies that the posteriors are also independent.

Assumptions

IRT makes a few strong assumptions:

First, observations are independent conditional on θ.
Two students' test answers are related only in so far that the students have similar ability.

Second, items are independent conditional on θ.
Getting one question right should not affect the probability of getting another question right, outside of the ability demonstrated in answering both questions.

Usually (but not necessarily), the priors $P(\theta_i)$ are assumed to be independent across observations. That implies that the posteriors are also independent.

Can you think of situations in which these assumptions are violated?

Other Uses of IRT

Psychometrics, used to measure latent self-esteem, depression, attachment anxiety.

Other Uses of IRT

Psychometrics, used to measure latent self-esteem, depression, attachment anxiety.

Computerized adaptive testing (Montgomery and Cutler 2013)

Other Uses of IRT

Psychometrics, used to measure latent self-esteem, depression, attachment anxiety.

Computerized adaptive testing (Montgomery and Cutler 2013)

Examples in political science:

- Cross-national variation in democracy (Treier and Jackman 2008)
- Ideal point estimates for:
  - members of Congress (Clinton, Jackman, and Rivers 2004)
  - Supreme Court Justices (Martin and Quinn 2002, Bailey and Maltzmann 2008)
  - state legislators (Shor and McCarty 2011)
  - member states in the UN (Voeten 2004)

Extensions of IRT

IRT is, in my opinion, terribly underutilized.

Not that is isn't used enough in research – but when it is used, it's used in too limited a way.

Extensions of IRT

IRT is, in my opinion, terribly underutilized.

Not that is isn't used enough in research – but when it is used, it's used in too limited a way.

The current practice is that only binary, or only ordinal, or only nominal items are used.

Extensions of IRT

IRT is, in my opinion, terribly underutilized.

Not that is isn't used enough in research – but when it is used, it's used in too limited a way.

The current practice is that only binary, or only ordinal, or only nominal items are used.

But all IRT is an extension of GLM. Anything we can do in GLM, we can do in IRT. Some extensions:

1. IRT when the items are ordinal, nominal

Extensions of IRT

IRT is, in my opinion, terribly underutilized.

Not that is isn't used enough in research – but when it is used, it's used in too limited a way.

The current practice is that only binary, or only ordinal, or only nominal items are used.

But all IRT is an extension of GLM. Anything we can do in GLM, we can do in IRT. Some extensions:

1. IRT when the items are ordinal, nominal
2. Count, proportion, continuous items, general test curves

Extensions of IRT

IRT is, in my opinion, terribly underutilized.

Not that is isn't used enough in research – but when it is used, it's used in too limited a way.

The current practice is that only binary, or only ordinal, or only nominal items are used.

But all IRT is an extension of GLM. Anything we can do in GLM, we can do in IRT. Some extensions:

1. IRT when the items are ordinal, nominal
2. Count, proportion, continuous items, general test curves
3. Multidimensional estimates of θ

Extensions of IRT

IRT is, in my opinion, terribly underutilized.

Not that is isn't used enough in research – but when it is used, it's used in too limited a way.

The current practice is that only binary, or only ordinal, or only nominal items are used.

But all IRT is an extension of GLM. Anything we can do in GLM, we can do in IRT. Some extensions:

1. IRT when the items are ordinal, nominal
2. Count, proportion, continuous items, general test curves
3. Multidimensional estimates of θ
4. Creating time dependent estimates of θ

Graded Response Model

The version of IRT that uses all ordinal items is called the **graded response model** (GRM).

Graded Response Model

The version of IRT that uses all ordinal items is called the **graded response model** (GRM).

The name comes from the idea that items aren't just correct or incorrect, but have varying degrees of correctness, with labels like A, B, C, D, F. The GRM uses this ordinal information.

Graded Response Model

The version of IRT that uses all ordinal items is called the **graded response model** (GRM).

The name comes from the idea that items aren't just correct or incorrect, but have varying degrees of correctness, with labels like A, B, C, D, F. The GRM uses this ordinal information.

Binary IRT is built on the logic of logistic regression. So, it makes sense that the GRM is built on top of a ordered logit model.

Graded Response Model

The GRM uses the same standard normal prior distributions on the values of the latent variable as binary logit:

$$\theta \sim N(0, 1)$$

Graded Response Model

The GRM uses the same standard normal prior distributions on the values of the latent variable as binary logit:

$$\theta \sim N(0, 1)$$

Also like binary IRT, GRM gets posterior estimates of each θ by multiplying the prior by every test curve. The question is: what should the test curves be?

Graded Response Model

The GRM uses the same standard normal prior distributions on the values of the latent variable as binary logit:

$$\theta \sim N(0, 1)$$

Also like binary IRT, GRM gets posterior estimates of each θ by multiplying the prior by every test curve. The question is: what should the test curves be?

Binary items can only be **0 or 1** so the two test curves are

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta-\beta)}} \quad \text{and} \quad P(X = 0) = 1 - \frac{1}{1 + e^{-\alpha(\theta-\beta)}}$$

where α is the item's discrimination and β is the item's difficulty.

Graded Response Model

But ordinal items can be equal to **many different ordered categories** (let's call the categories 1, 2, ..., K). So we need K test curves. The first category's curve is:

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

Graded Response Model

But ordinal items can be equal to **many different ordered categories** (let's call the categories 1, 2, ..., K). So we need K test curves. The first category's curve is:

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

the test curve for categories 2 through $K - 1$ are

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

Graded Response Model

But ordinal items can be equal to **many different ordered categories** (let's call the categories 1, 2, ..., K). So we need K test curves. The first category's curve is:

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

the test curve for categories 2 through $K - 1$ are

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

and the test curve for the last category is

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

Graded Response Model

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

Graded Response Model

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

These are the exact same functions as the link function for ordered logit, only the linear model is rearranged to produce difficulty and discrimination parameters.

Graded Response Model

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

These are the exact same functions as the link function for ordered logit, only the linear model is rearranged to produce difficulty and discrimination parameters.

There is **one discrimination parameter** α for the item, but $K - 1$ difficulty parameters for the K categories. Why?

Graded Response Model

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

These are the exact same functions as the link function for ordered logit, only the linear model is rearranged to produce difficulty and discrimination parameters.

There is **one discrimination parameter** α for the item, but $K - 1$ difficulty parameters for the K categories. Why? **Because these difficulty parameters take the place of the ordered logit cutpoints**.

Graded Response Model

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

Graphically, the first and last of the ordinal test curves are S-shaped, just like the binary IRT test curves.

Graded Response Model

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

But the intermediate test curves look like bell curves. These are logistic, not normal, bell curves, but are very similar.

Graded Response Model

$$P(X = 1) = \frac{1}{1 + e^{-\alpha(\theta - \beta_1)}},$$

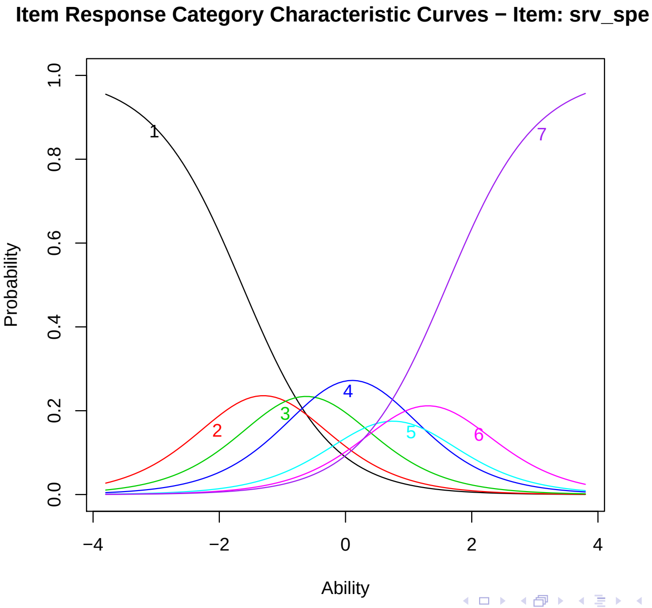
$$P(X = j) = \frac{1}{1 + e^{-\alpha(\theta - \beta_j)}} - \frac{1}{1 + e^{-\alpha(\theta - \beta_{j-1})}},$$

$$P(X = K) = 1 - \frac{1}{1 + e^{-\alpha(\theta - \beta_K)}}.$$

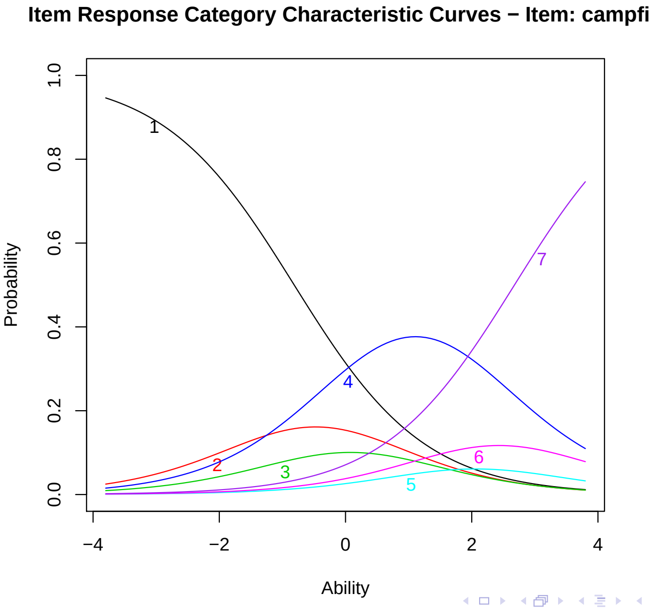
But the intermediate test curves look like bell curves. These are logistic, not normal, bell curves, but are very similar.

Either way, the curves represent the probability that an observation with a particular θ responds with each category. If you plot all the curves together and draw any vertical line, the y-values (probabilities) add to 1.

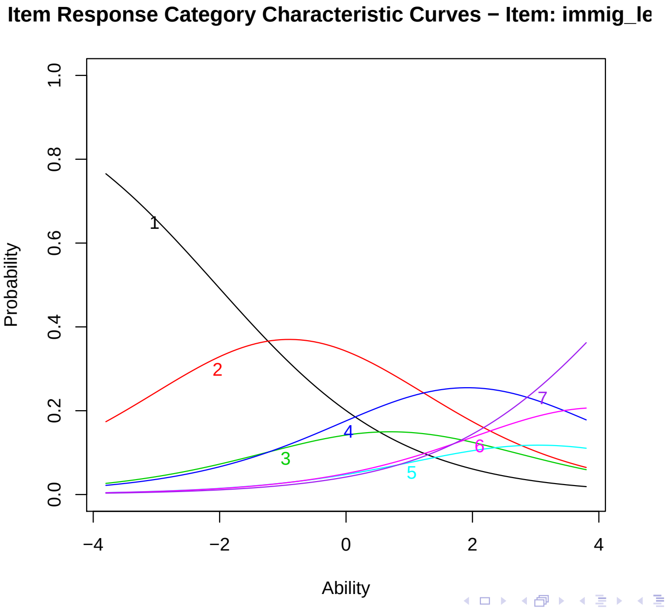
Graded Response Model



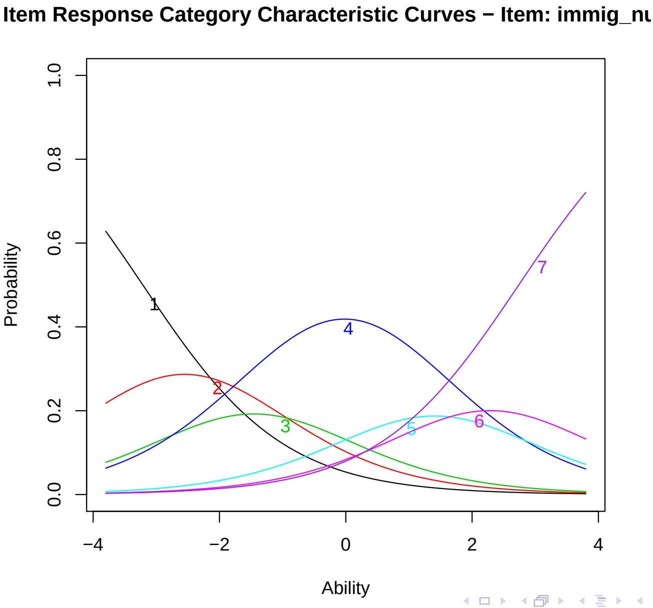
Graded Response Model



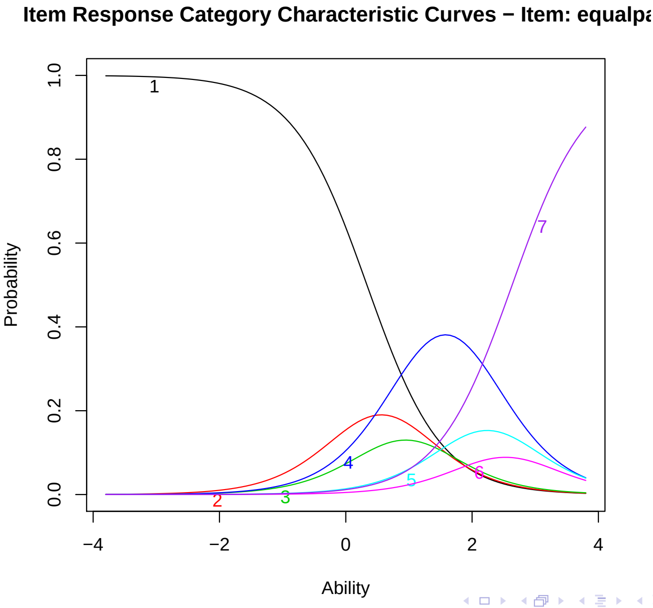
Graded Response Model



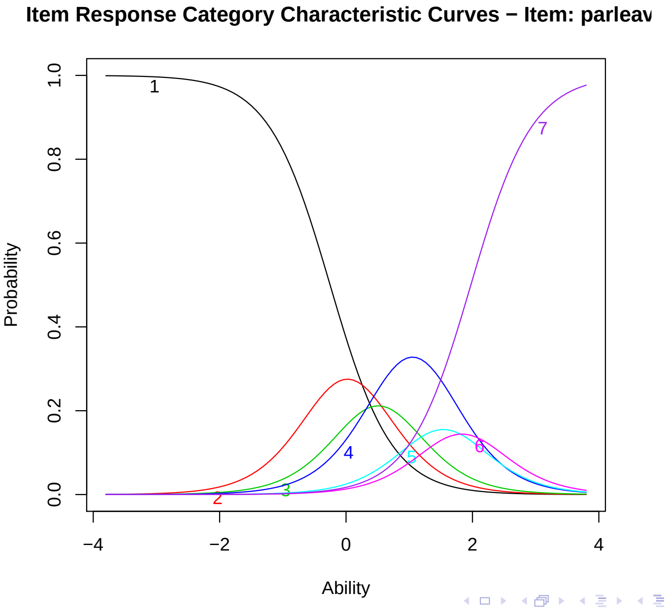
Graded Response Model



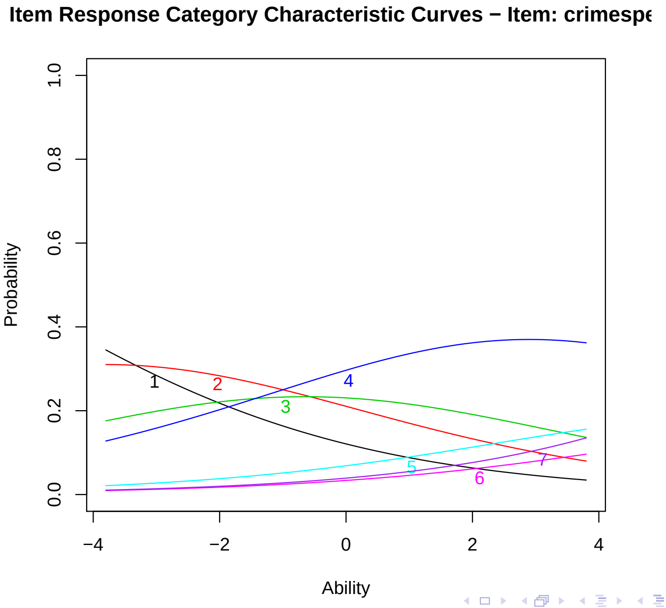
Graded Response Model



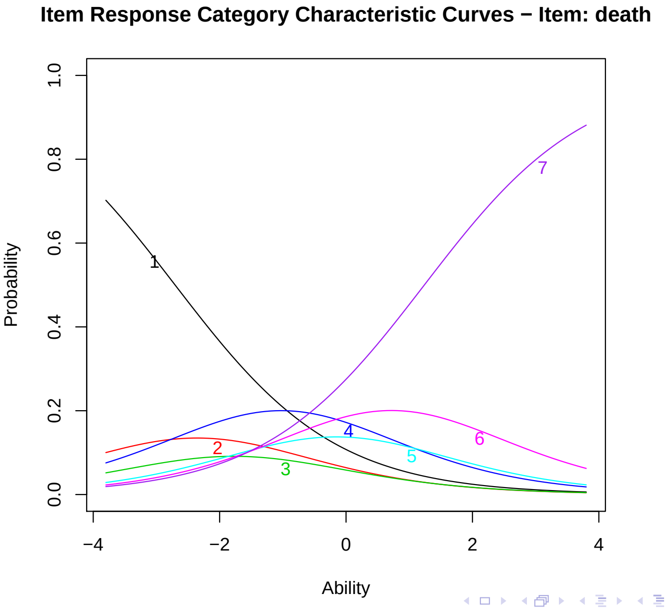
Graded Response Model



Graded Response Model



Graded Response Model



Nominal IRT

It is possible to work with unordered categorical items as well. These items are much more rare on a test, but common in political data. Some examples:

- vote choices,
- regime types,
- conflict outcomes,
- demographics like race, religion, marital status.

Nominal IRT

It is possible to work with unordered categorical items as well. These items are much more rare on a test, but common in political data. Some examples:

- vote choices,
- regime types,
- conflict outcomes,
- demographics like race, religion, marital status.

The nominal IRT model is built upon **multinomial logit**. Consider an item with 3 categories. The test curves are:

Nominal IRT

It is possible to work with unordered categorical items as well. These items are much more rare on a test, but common in political data. Some examples:

- vote choices,
- regime types,
- conflict outcomes,
- demographics like race, religion, marital status.

The nominal IRT model is built upon **multinomial logit**. Consider an item with 3 categories. The test curves are:

$$P(X = 1) = \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}},$$

Nominal IRT

It is possible to work with unordered categorical items as well. These items are much more rare on a test, but common in political data. Some examples:

- vote choices,
- regime types,
- conflict outcomes,
- demographics like race, religion, marital status.

The nominal IRT model is built upon **multinomial logit**. Consider an item with 3 categories. The test curves are:

$$P(X = 1) = \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}},$$

$$P(X = 2) = \frac{1}{1 + e^{-\alpha_2(\theta - \beta_2)}},$$

Nominal IRT

It is possible to work with unordered categorical items as well. These items are much more rare on a test, but common in political data. Some examples:

- vote choices,
- regime types,
- conflict outcomes,
- demographics like race, religion, marital status.

The nominal IRT model is built upon **multinomial logit**. Consider an item with 3 categories. The test curves are:

$$P(X=1)=\frac{1}{1+e^{-\alpha_1(\theta-\beta_1)}},$$

$$P(X=2)=\frac{1}{1+e^{-\alpha_2(\theta-\beta_2)}},$$

$$P(X=3)=1-\frac{1}{1+e^{-\alpha_1(\theta-\beta_1)}}-\frac{1}{1+e^{-\alpha_2(\theta-\beta_2)}}.$$

Nominal IRT

$$
\begin{aligned}
P(X = 1) &= \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}}, \\
P(X = 2) &= \frac{1}{1 + e^{-\alpha_2(\theta - \beta_2)}}, \\
P(X = 3) &= 1 - \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}} - \frac{1}{1 + e^{-\alpha_2(\theta - \beta_2)}}.
\end{aligned}
$$

Nominal IRT

$$P(X = 1) = \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}},$$

$$P(X = 2) = \frac{1}{1 + e^{-\alpha_2(\theta - \beta_2)}},$$

$$P(X = 3) = 1 - \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}} - \frac{1}{1 + e^{-\alpha_2(\theta - \beta_2)}}.$$

Note that unlike the GRM, we now have **different discrimination parameters** for every category except one.

Nominal IRT

$$P(X = 1) = \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}},$$

$$P(X = 2) = \frac{1}{1 + e^{-\alpha_2(\theta - \beta_2)}},$$

$$P(X = 3) = 1 - \frac{1}{1 + e^{-\alpha_1(\theta - \beta_1)}} - \frac{1}{1 + e^{-\alpha_2(\theta - \beta_2)}}.$$

Note that unlike the GRM, we now have **different discrimination parameters** for every category except one.

In this case, the discrimination and difficulty parameters are interpreted relative to the base category.

Other kinds of IRT

There is absolutely nothing stopping us from applying the GLM logic to items of all other kinds.

Other kinds of IRT

There is absolutely nothing stopping us from applying the GLM logic to items of all other kinds.

Consider **count items**. What can we do to create a count IRT model?

Other kinds of IRT

There is absolutely nothing stopping us from applying the GLM logic to items of all other kinds.

Consider **count items**. What can we do to create a count IRT model? Use a negative binomial test curve:

$$P(X = c) = \binom{c + r + 1}{c} (1 - p)^r p^c,$$

where r is the negative binomial overdispersion parameter, and

Other kinds of IRT

There is absolutely nothing stopping us from applying the GLM logic to items of all other kinds.

Consider **count items**. What can we do to create a count IRT model? Use a negative binomial test curve:

$$P(X = c) = \binom{c + r + 1}{c} (1 - p)^r p^c,$$

where r is the negative binomial overdispersion parameter, and

$$p = \frac{1}{1 + e^{-\alpha(\theta - \beta)}}.$$

Other kinds of IRT

There is absolutely nothing stopping us from applying the GLM logic to items of all other kinds.

Consider **count items**. What can we do to create a count IRT model? Use a negative binomial test curve:

$$P(X = c) = \binom{c + r + 1}{c}(1 - p)^r p^c,$$

where r is the negative binomial overdispersion parameter, and

$$p = \frac{1}{1 + e^{-\alpha(\theta - \beta)}}.$$

Likewise, we can build an IRT model from any GLM: normal (for continuous), beta (for proportions), Weibull (for durations), gamma (for non-negative continuous), etc.

Other kinds of IRT

There's also nothing stopping us from putting all of these items together in one big IRT model. All we have to do it multiply the prior by every item's unique test curve.

Other kinds of IRT

There's also nothing stopping us from putting all of these items together in one big IRT model. All we have to do it multiply the prior by every item's unique test curve.

This model is the cutting edge of measurement statistics. It is **flexible** because it can handle all sorts of model specifications and item types. But it is also built on top of theoretically driven GLMs.

Other kinds of IRT

There's also nothing stopping us from putting all of these items together in one big IRT model. All we have to do it multiply the prior by every item's unique test curve.

This model is the cutting edge of measurement statistics. It is **flexible** because it can handle all sorts of model specifications and item types. But it is also built on top of theoretically driven GLMs.

The best new examples of clever measurement almost all start with IRT and customize it for a specific application by using alternative GLMs or a game theoretic model for test curves (as DW-NOMINATE does).

Other kinds of IRT

There's also nothing stopping us from putting all of these items together in one big IRT model. All we have to do it multiply the prior by every item's unique test curve.

This model is the cutting edge of measurement statistics. It is **flexible** because it can handle all sorts of model specifications and item types. But it is also built on top of theoretically driven GLMs.

The best new examples of clever measurement almost all start with IRT and customize it for a specific application by using alternative GLMs or a game theoretic model for test curves (as DW-NOMINATE does).

The limiting factor: making the computer do what we want. We will use a powerful tool for exactly that, called Stan, next week.

Something I Did: Time-Series Item Response Theory (TSIRT)

Cases: $N = 1$

Timepoints: $T \to \infty$

Data: X is $(T \times K)$, T timepoints and K items.
Items may be categorical, continuous, count, or proportion.

Latent variable: θ_t, unidimensional, derived from shared covariance of columns of X

Prior: Integrated time series (also used by Martin and Quinn (2002))

$$\theta_0 \sim N(0, \sigma^2), \qquad \theta_t \sim N(\theta_{t-1}, \sigma^2),$$

where $t \in \{1, 2, \ldots, T\}$, and σ^2 is fixed across t and estimated.

Test Curves

$$p_{tj} = \frac{1}{1 + \exp(-\alpha_j(\theta_t - \beta_j))}$$

Test Curves

$$p_{tj} = \frac{1}{1 + \exp(-\alpha_j(\theta_t - \beta_j))}$$

Binary items: $X_j \sim \text{Bernoulli}(p_{tj})$,

Count items: $X_j \sim NB(p_{tj}, r_j)$,

Proportion items: $X_j \sim \text{Beta}(p_{tj}, \phi_j)$,

Standardized continuous items: $X_j \sim N(\theta_t, \alpha_j^2)$, Test Curve: the distribution of an item conditional on θ_t and on the item parameters.

Binary items: $X_j \sim \text{Bernoulli}(p)$, where

$$p_{tj} = \frac{1}{1 + \exp(-\alpha_j(\theta_t - \beta_j))}.$$

Item parameters to estimate:

- α_j – discrimination
- β_j – difficulty

Standardized continuous items: $X_j \sim N(\theta_t, \alpha_j^2)$.
Item parameter to estimate:

- α_j – standard deviation (discrimination)

Count items: X_j distributed Negative Binomial:

$$f(X_j|\theta_t, \alpha_j, \beta_j, r_j) = \begin{pmatrix} X_j + r_j - 1 \\ X_j \end{pmatrix} (1 - p_{tj})^r {p_{tj}}^{X_j},$$

$$p_{tj} = \frac{1}{1 + \exp(-\alpha_j(\theta_t - \beta_j))}.$$

Item parameters to estimate:

- α_j – discrimination
- β_j – difficulty
- r_j – the number of negative draws before the experiment is terminated

Proportion items: $X_j \in [0, 1]$, distributed Beta:

$$f(X_j|\theta_t, \alpha_j, \beta_j, \phi_j) = \frac{(X_j)^{p_{tj}\phi_j - 1}(1 - X_j)^{(1-p_{tj})\phi_j - 1}}{B\Big(p_{tj}\phi_j, (1 - p_{tj})\phi_j\Big)},$$

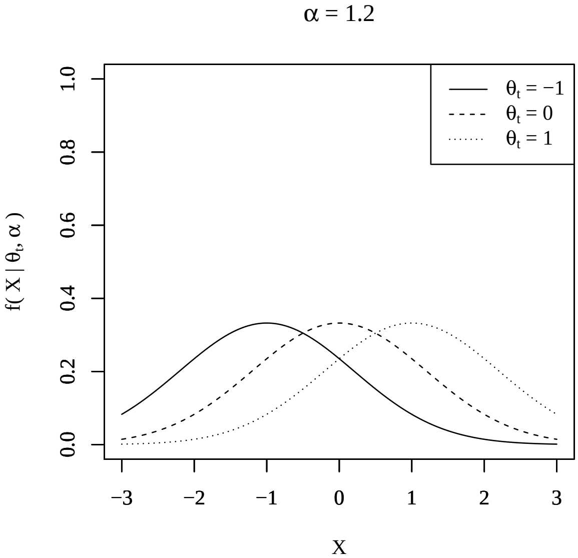
$$p_{tj} = \frac{1}{1 + \exp(-\alpha_j(\theta_t - \beta_j))},$$

where $B()$ is the Beta function.

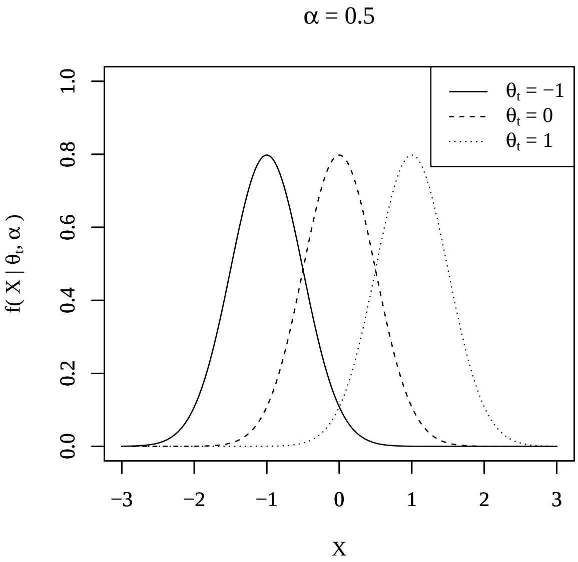
Item parameters to estimate:

- α_j – discrimination
- β_j – difficulty
- ϕ_j – total count parameter

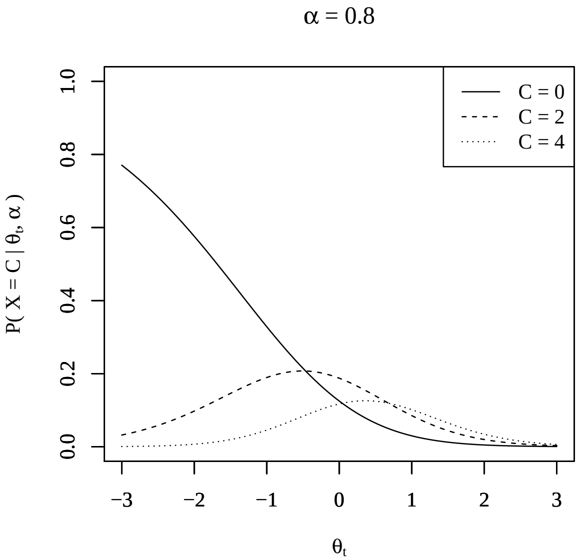
Normal test curve, low discrimination



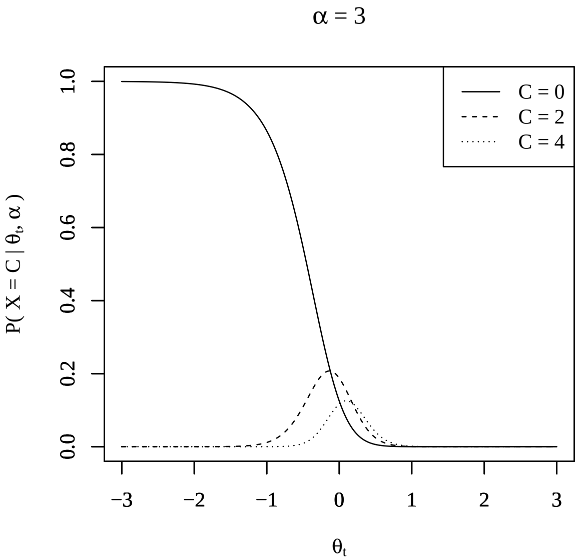
Normal test curve, high discrimination



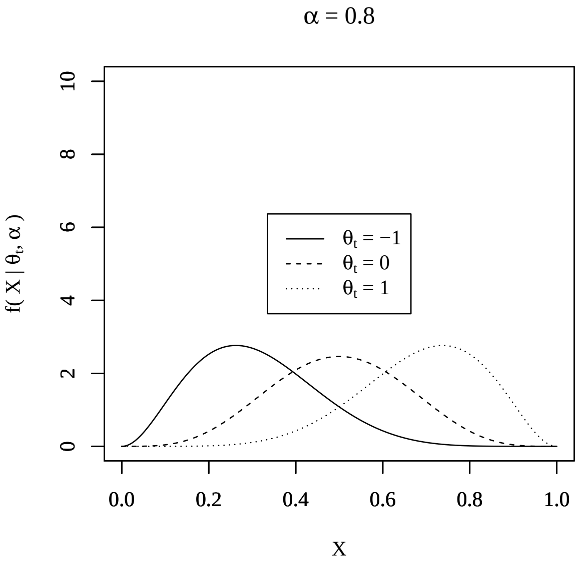
Count test curve, low discrimination



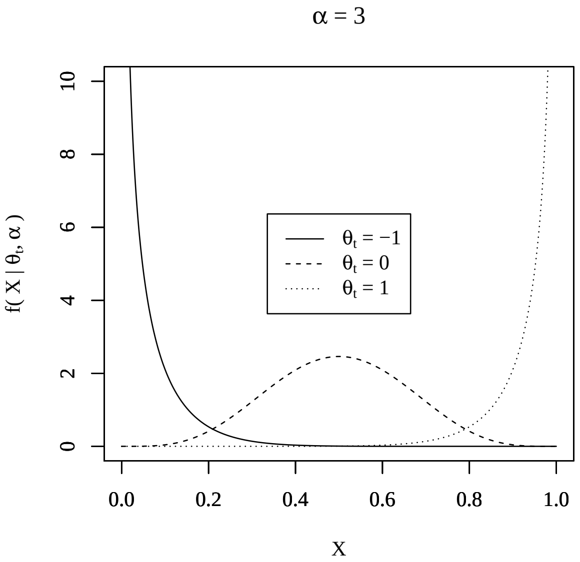
Count test curve, high discrimination



Proportion test curve, low discrimination



Proportion test curve, high discrimination



TSIRT is implemented as a fully Bayesian model, and θ_t, σ^2, and the item parameters are estimated through MCMC:

$$\begin{aligned}
P(\theta_t, \sigma^2, \alpha, \beta, r, \phi | X) \quad &\propto \quad P_\theta(\theta_t) \cdot P_{\sigma^2}(\sigma^2) \cdot P_\alpha(\alpha) \cdot P_\beta(\beta) \cdot P_r(r) \cdot P_\phi(\phi) \\
\text{(binary)} \quad & \quad \times \prod_{k=1}^{K_B} f_{Bk}(X|\theta_t, \alpha_k, \beta_k) \\
\text{(count)} \quad & \quad \times \prod_{k=1}^{K_C} f_{Ck}(X|\theta_t, \alpha_k, \beta_k, r_k) \\
\text{(proportion)} \quad & \quad \times \prod_{k=1}^{K_P} f_{Pk}(X|\theta_t, \alpha_k, \beta_k, \phi_k). \\
\text{(continuous)} \quad & \quad \times \prod_{k=1}^{K_N} f_{Nk}(X|\theta_t, \alpha_k)
\end{aligned}$$

Convergence assessed through multiple chains and $\hat{R}$ statistic (Gelman and Rubin 1992).

TSIRT is implemented as a fully Bayesian model, and θ_t, σ^2, and the item parameters are estimated through MCMC:

$$
\begin{aligned}
P(\theta_t, \sigma^2, \alpha, \beta, r, \phi | X) \quad & \propto \quad P_\theta(\theta_t) \cdot P_{\sigma^2}(\sigma^2) \cdot P_\alpha(\alpha) \cdot P_\beta(\beta) \cdot P_r(r) \cdot P_\phi(\phi) \\
\text{(binary)} \quad & \quad \times \prod_{k=1}^{K_B} f_{Bk}(X|\theta_t, \alpha_k, \beta_k) \\
\text{(count)} \quad & \quad \times \prod_{k=1}^{K_C} f_{Ck}(X|\theta_t, \alpha_k, \beta_k, r_k) \\
\text{(proportion)} \quad & \quad \times \prod_{k=1}^{K_P} f_{Pk}(X|\theta_t, \alpha_k, \beta_k, \phi_k). \\
\text{(continuous)} \quad & \quad \times \prod_{k=1}^{K_N} f_{Nk}(X|\theta_t, \alpha_k)
\end{aligned}
$$

Convergence assessed through multiple chains and $\hat{R}$ statistic (Gelman and Rubin 1992).

Posterior estimates of θ have serial dependence because the prior $P_\theta(\theta)$ has serial dependence.

Example: the Israeli/Palestinian Conflict, 1971-2013

Spoiler Violence (Kydd & Walter 2002)

- Violence surrounding cooperation aimed at undermining talks
- Excluded factions aim to spoil peace
- Occurs during talks and implementation
- Short term

Bueno de Mesquita (2005)

- Moderates are pulled into cooperation leaving extremists in opposition
- Increased militancy leads to higher violence
- Sustained increase in violence following negotiations
- Long term

Example: the Israeli/Palestinian Conflict, 1971-2013

Spoiler Violence (Kydd & Walter 2002)

- Violence surrounding cooperation aimed at undermining talks
- Excluded factions aim to spoil peace
- Occurs during talks and implementation
- Short term

Bueno de Mesquita (2005)

- Moderates are pulled into cooperation leaving extremists in opposition
- Increased militancy leads to higher violence
- Sustained increase in violence following negotiations
- Long term

Data: Dyadic event counts via GDELT (Leetaru and Schrodt 2013), compiled quarterly, 1971-2012.

Cooperation Data

Event	Direction	Mean	SD	Min	Max
Provide Aid	ISR → PAL	28.7	16.5	0	96
	PAL → ISR	17.5	11.2	0	59
Appeal for Cooperation	ISR → PAL	76.4	33.4	33	206
	PAL → ISR	77.7	30.2	9	184
Cooperative Action	ISR → PAL	60.1	25.1	15	163
	PAL → ISR	72.1	28.4	0	182
Express Intent to Cooperate	ISR → PAL	143.1	56.4	35	269
	PAL → ISR	137.4	50.3	18	277
Optimistic Statement	ISR → PAL	42.4	20.4	0	102
	PAL → ISR	42.7	18.8	0	105
Release Prisoners	ISR → PAL	13.2	13.9	0	86
	PAL → ISR	24.7	20.1	0	111
Concessions	ISR → PAL	44.2	19.9	0	103
	PAL → ISR	33.9	18.9	0	95
Formal Agreement		20.5	19.9	1	104
Meet		115.1	47.9	31	260
Negotiate		58.9	32.0	7	152

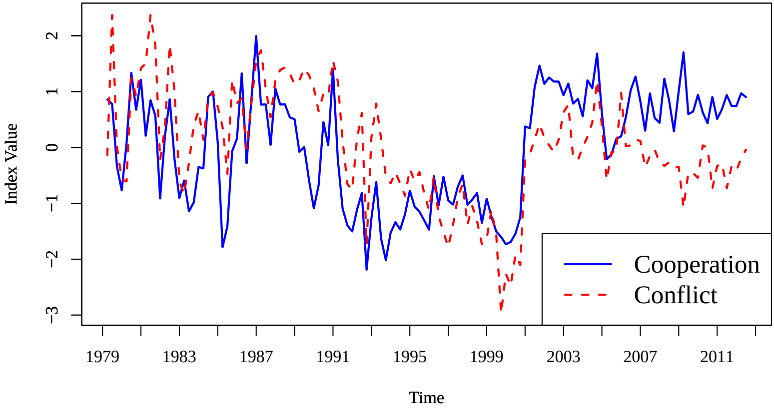
Conflict Data

Event	Direction	Mean	SD	Min	Max
Administrative Sanctions	ISR $\rightarrow$ PAL	17.8	10.6	0	61
	PAL $\rightarrow$ ISR	33.0	16.9	0	94
Assassination Attempts	ISR $\rightarrow$ PAL	9.9	10.7	0	66
	PAL $\rightarrow$ ISR	10.0	11.0	0	53
Coercion	ISR $\rightarrow$ PAL	3.0	3.8	0	23
	PAL $\rightarrow$ ISR	0.8	1.7	0	10
Denounce	ISR $\rightarrow$ PAL	48.5	22.6	0	131
	PAL $\rightarrow$ ISR	52.1	25.0	0	194
Deportation	ISR $\rightarrow$ PAL	6.6	7.7	2	55
	PAL $\rightarrow$ ISR	5.1	6.0	1	39
Detention	ISR $\rightarrow$ PAL	39.6	27.0	0	137
	PAL $\rightarrow$ ISR	50.1	25.2	0	167
Embargo	ISR $\rightarrow$ PAL	4.0	6.4	0	34
	PAL $\rightarrow$ ISR	4.6	5.5	0	25
Mass Killing	ISR $\rightarrow$ PAL	4.2	5.0	1	25
	PAL $\rightarrow$ ISR	5.1	6.7	0	28

Conflict Data

Event	Direction	Mean	SD	Min	Max
Conventional Military Action	ISR → PAL	135.6	69.5	38	446
	PAL → ISR	132.6	62.5	39	396
Occupation	ISR → PAL	37.3	21.6	0	104
	PAL → ISR	18.1	14.4	0	110
Action Against Property	ISR → PAL	21.7	16.6	0	78
	PAL → ISR	10.9	9.5	0	50
Restrict Movement	ISR → PAL	7.0	7.9	0	57
	PAL → ISR	9.9	10.4	0	74
Threaten	ISR → PAL	62.3	24.2	7	144
	PAL → ISR	61.3	24.4	0	119
Unconventional Violence	ISR → PAL	53.3	27.9	0	175
	PAL → ISR	60.6	26.8	0	142
Civil Unrest	ISR → PAL	22.0	15.7	0	86
	PAL → ISR	30.8	22.3	1	141
Violent Repression	ISR → PAL	1.7	2.7	0	14
	PAL → ISR	3.7	4.8	0	29

Cooperation and Conflict Indices

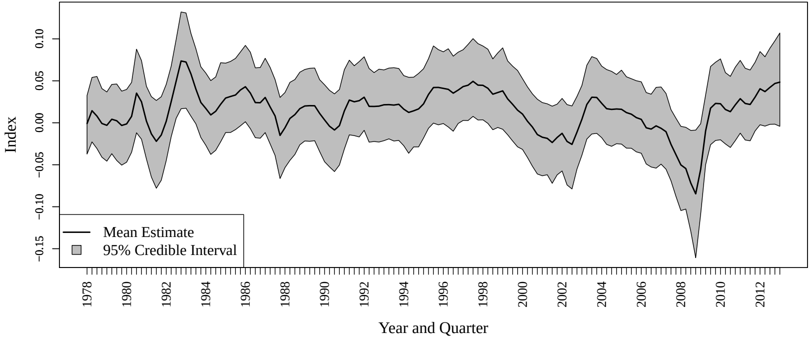


U.S. Economic Performance Since 1978

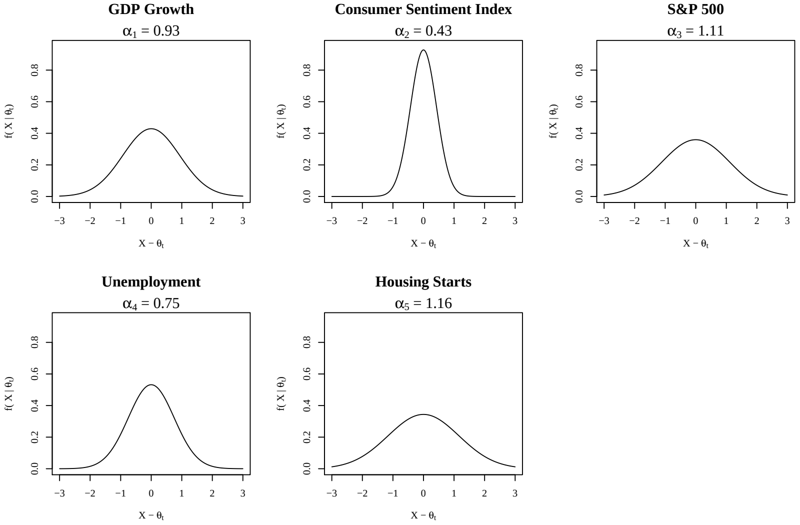
What economic indicator is the best measure of the overall performance of the economy?

Table: Indicators of U.S. Quarterly Economic Performance, 1978-2013.

Indicator	Mean	Best	Worst
GDP Growth	2.71	16.7 (1978, Q2)	-8.9 (2008, Q4)
Consumer Sentiment Index	85.3	110.1 (2000, Q1)	51.1 (1980, Q2)
S&P 500, % Change	2.20	20.2 (1982, Q4)	-27.2 (2008, Q4)
Unemployment Rate	6.42	3.9 (2000, Q4)	10.7 (1982, Q4)
Housing Starts, % Change	-0.18	31.5 (1980, Q3)	-23.1 (2008, Q4)



Captures the the recessions of the early 1980s, the stock market crash of 1987, the recession of the early 1990s, the burst of the "dot-com" bubble in the early 2000s, and the recession of 2008.


GDP Growth
α1 = 0.93
Consumer Sentiment Index
α2 = 0.43
S&P 500
α3 = 1.11
Unemployment
α4 = 0.75
Housing Starts
α5 = 1.16
f(X | θt)
X − θt