Survey Research and Design

Generating Scales to Measure Latent Traits

William Marble

November 14, 2023

Measuring Latent Traits

Latent traits are stable (psychological?) dispositions that influence more immediate attitudes and behavior.

Examples:

- ideology
- scholastic “aptitude”
- risk aversion
- personality traits

Last Time

Drawbacks of direct elicitation:

- leaves definition of the construct up to the respondent
- different interepretations make interpersonal comparisons difficult or impossible
- may suffer from lack of content validity or convergent validity

The Latent Variable View

Latent traits influence more immediate attitudes and behavior.

The Latent Variable View

Latent traits influence more immediate attitudes and behavior.

Suggests a different research strategy:

- ask several questions that are thought to be influenced by the latent trait
- responses should give us some information about the respondents value on the latent trait
- combining responses from several together should give us a better measure than any one question – which may be imperfectly related to the trait of interest

The Latent Variable View

Latent traits influence more immediate attitudes and behavior.

Suggests a different research strategy:

- ask several questions that are thought to be influenced by the latent trait
- responses should give us some information about the respondents value on the latent trait
- combining responses from several together should give us a better measure than any one question – which may be imperfectly related to the trait of interest

Research design considerations:

- how many questions to ask?
- which questions to ask?
- how to combine questions?

Educational Testing Analogy

Latent traits are familiar: educational tests are used to measure latent traits!

Educational Testing Analogy

Latent traits are familiar: educational tests are used to measure latent traits!

Math SAT as an example. The test is designed to measure “understanding of mathematical concepts, procedural skill and fluency in math, and ability to apply those concepts and skills to real-world problems.”

Educational Testing Analogy

Latent traits are familiar: educational tests are used to measure latent traits!

Math SAT as an example. The test is designed to measure “understanding of mathematical concepts, procedural skill and fluency in math, and ability to apply those concepts and skills to real-world problems.”

Key features of the SAT math section (for our purposes):

- There is a right answer to each question
- The questions are tailored to measure mathematical aptitude (assume this is true for now)
- There are about 70 questions total

Educational Testing Analogy

Latent traits are familiar: educational tests are used to measure latent traits!

Math SAT as an example. The test is designed to measure "understanding of mathematical concepts, procedural skill and fluency in math, and ability to apply those concepts and skills to real-world problems."

Key features of the SAT math section (for our purposes):

- There is a right answer to each question
- The questions are tailored to measure mathematical aptitude (assume this is true for now)
- There are about 70 questions total

Question: How to determine an individual student's score?

Data Structure and Notation

Suppose we collect N students' responses to each of J questions. We observe y_{ij} which indicates whether respondent i got question j right ($y_{ij} = 1$) or wrong ($y_{ij} = 0$). The data is in tabular form:

Student (i)	y_{i1}	y_{i2}	y_{i3}	$\dots$	y_{iJ}	Total Score
1	1	1	0	$\dots$	1	?
2	0	1	1	$\dots$	1	?
3	0	1	1	$\dots$	0	?
4	1	1	0	$\dots$	0	?
5	1	1	0	$\dots$	1	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
N	1	1	1	$\dots$	0	?

Mapping Math Aptitude to Ideology

How does the SAT relate to public opinion?

Mapping Math Aptitude to Ideology

How does the SAT relate to public opinion?

Similarities

- Mathematical aptitude $\approx$ ideology
- Math questions $\approx$ questions on policy attitudes (agree/disagree)

Main Difference

- There is no "right answer" to policy questions $\rightsquigarrow$ we'll come back to this

Cultural Ideology Scale, from Last Time

- *Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if. . . [6 scenarios]*
- *There's been a lot of discussion about the way morals and attitudes about sex are changing in this country. If two adults of the same sex have sexual relations, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?*
- *Do you strongly agree, agree, disagree, or strongly disagree that methods of birth control should be available to teenagers between the ages of 14 and 16 if their parents do not approve?*
- *When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient's life by some painless means if the patient and his family request it?*
- *The United States Supreme Court has ruled that no state or local government may require the reading of the Lord's Prayer or Bible verses in public schools. What are your views on this — do you approve or disapprove of the court ruling?*
- *Would you be for or against sex education in the public schools?*

The Number of Items

Why does the SAT ask 70 questions? Why not just ask 1? Why not ask just 5?

The Number of Items

Why does the SAT ask 70 questions? Why not just ask 1? Why not ask just 5?

- More questions $\rightsquigarrow$ more information about each individual respondent
- Each question is influenced by the construct, but also influenced by other factors (random error, small differences in emphasis in class, etc.)
- The hope: combining many questions into one score minimizes the impact of other factors on the score
- Same principle justifies combining several survey items into a scale, rather than relying on a single question

The Number of Items

Why does the SAT ask 70 questions? Why not just ask 1? Why not ask just 5?

- More questions $\rightsquigarrow$ more information about each individual respondent
- Each question is influenced by the construct, but also influenced by other factors (random error, small differences in emphasis in class, etc.)
- The hope: combining many questions into one score minimizes the impact of other factors on the score
- Same principle justifies combining several survey items into a scale, rather than relying on a single question

The trade-off for survey research: asking too many questions leads to survey fatigue and takes up space that could be used for other questions.

$\rightsquigarrow$ How to pick questions?

Underlying Principles

If all questions are measuring the same construct, we should expect that people who get one question right are more likely to get the other questions right.

In notation, we expect answers to question j and j' to be *positively correlated*, but not perfectly so:

$$0 < \text{Cor}(y_{ij}, y_{ij'}) < 1$$

Underlying Principles

If all questions are measuring the same construct, we should expect that people who get one question right are more likely to get the other questions right.

In notation, we expect answers to question j and j' to be *positively correlated*, but not perfectly so:

$$0 < \text{Cor}(y_{ij}, y_{ij'}) < 1$$

Suggests that questions that aren't highly correlated with other questions might not be good measures of the underlying construct.

Example Test Data

We can examine the correlation matrix, which shows the correlations between each column in the data frame:

	y1	y2	y3	y4	y5
y1		0.01	0.03	0.02	0.08
y2	0.01		0.15	0.20	0.10
y3	0.03	0.15		0.18	0.21
y4	0.02	0.20	0.18		0.14
y5	0.08	0.10	0.21	0.14	

Example Test Data

We can examine the correlation matrix, which shows the correlations between each column in the data frame:

	y1	y2	y3	y4	y5
y1		0.01	0.03	0.02	0.08
y2	0.01		0.15	0.20	0.10
y3	0.03	0.15		0.18	0.21
y4	0.02	0.20	0.18		0.14
y5	0.08	0.10	0.21	0.14	

Question 1 is not very correlated with other questions $\rightsquigarrow$ maybe not a good question.

Assessing Scale Consistency: Cronbach's Alpha

To measure whether items "go together," a simple and commonly used measure is Cronbach's alpha:

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

N : number of items

$\bar{c}$: average inter-item covariance across items

$\bar{v}$ average item variance

Assessing Scale Consistency: Cronbach's Alpha

To measure whether items "go together," a simple and commonly used measure is Cronbach's alpha:

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

N : number of items

$\bar{c}$: average inter-item covariance across items

$\bar{v}$ average item variance

Ranges from 0 (no correlation between items) to 1 (perfect correlation).

Rule-of-thumb cutoff: reliable/internally consistent scales have $\alpha \geq 0.7$

In R: `alpha()` function from `psych` package

How Should We Assign Scores?

How Should We Assign Scores?

- Raw proportion of questions correct

How Should We Assign Scores?

- Raw proportion of questions correct
- Weight some questions more highly than others

How Should We Assign Scores?

- Raw proportion of questions correct
- Weight some questions more highly than others(how?)
- Average together standardized scores for each question

How Should We Assign Scores?

- Raw proportion of questions correct
- Weight some questions more highly than others(how?)
- Average together standardized scores for each question
- Use IRT model

Percent of Right Answers

The score could be the proportion of questions the student got right:

$$\begin{aligned}\text{RawScore}_i &= \frac{1}{J}\left[y_{i1} + y_{i2} + y_{i3} + \cdots + y_{iJ}\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} y_{ij}\end{aligned}$$

Advantages

Percent of Right Answers

The score could be the proportion of questions the student got right:

$$\begin{aligned}\text{RawScore}_i &= \frac{1}{J}\left[y_{i1} + y_{i2} + y_{i3} + \cdots + y_{iJ}\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} y_{ij}\end{aligned}$$

Advantages

- Simple to compute, easy to interpret
- Works well if all the questions are about equally as good at measuring underlying construct

Percent of Right Answers

The score could be the proportion of questions the student got right:

$$\begin{aligned}\text{RawScore}_i &= \frac{1}{J}\left[y_{i1} + y_{i2} + y_{i3} + \cdots + y_{iJ}\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} y_{ij}\end{aligned}$$

Advantages

- Simple to compute, easy to interpret
- Works well if all the questions are about equally as good at measuring underlying construct

Disadvantage

Percent of Right Answers

The score could be the proportion of questions the student got right:

$$\begin{aligned} \text{RawScore}_i &= \frac{1}{J}\left[y_{i1} + y_{i2} + y_{i3} + \cdots + y_{iJ}\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} y_{ij} \end{aligned}$$

Advantages

- Simple to compute, easy to interpret
- Works well if all the questions are about equally as good at measuring underlying construct

Disadvantage

- Assumes that all questions are equally informative

Not All Questions Are Equal

What if almost everyone gets a question right or wrong?

Student (i)	y_{i1}	y_{i2}	y_{i3}	$\dots$	y_{iJ}	Total Score
1	1	1	0	$\dots$	1	?
2	0	1	0	$\dots$	1	?
3	0	1	1	$\dots$	0	?
4	1	1	0	$\dots$	0	?
5	1	1	0	$\dots$	1	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
N	1	1	1	$\dots$	0	?

Not All Questions Are Equal

What if almost everyone gets a question right or wrong?

Student (i)	y_{i1}	y_{i2}	y_{i3}	$\dots$	y_{iJ}	Total Score
1	1	1	0	$\dots$	1	?
2	0	1	0	$\dots$	1	?
3	0	1	1	$\dots$	0	?
4	1	1	0	$\dots$	0	?
5	1	1	0	$\dots$	1	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
N	1	1	1	$\dots$	0	?

Question 2: very easy, doesn't tell us much about aptitude

Question 3: pretty hard, should tell us a lot about aptitude

Averaging Standardized Scores

Instead of taking the proportion of right answers, we can **standardize** the columns first. Transform the columns into *z-scores*: make all columns have mean 0 and standard deviation 1.

$$z_{ij} = \frac{y_{ij} - \bar{y}_j}{\mathrm{sd}(y_j)}$$

Averaging Standardized Scores

Instead of taking the proportion of right answers, we can **standardize** the columns first. Transform the columns into *z-scores*: make all columns have mean 0 and standard deviation 1.

$$z_{ij} = \frac{y_{ij} - \bar{y}_j}{\text{sd}(y_j)}$$

Then take the average *z*-score across columns:

$$\begin{aligned}\text{StandardizedScore}_i &= \frac{1}{J}\left[z_{i1} + z_{i2} + z_{i3} + \cdots + z_{iJ}\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} z_{ij}\end{aligned}$$

Averaging Standardized Scores

Instead of taking the proportion of right answers, we can **standardize** the columns first. Transform the columns into *z-scores*: make all columns have mean 0 and standard deviation 1.

$$z_{ij} = \frac{y_{ij} - \bar{y}_j}{\text{sd}(y_j)}$$

Then take the average z-score across columns:

$$\begin{aligned}\text{StandardizedScore}_i &= \frac{1}{J}\left[z_{i1} + z_{i2} + z_{i3} + \cdots + z_{iJ}\right] \\ &= \frac{1}{J}\sum_{j=1}^{J} z_{ij}\end{aligned}$$

Properties:

- questions with lower variation $\rightsquigarrow$ higher weight (smaller SD in denominator)
- penalized more for getting easy questions wrong and rewarded more for getting harder questions right

Problems with Standarized Score

- A question could be hard and also unrelated to the construct of interest
- E.g., if the math SAT included a question about Roman history $\rightsquigarrow$ might be hard, but it's not informative about math aptitude
- Nonetheless the standardized scale would give that item a high weight

IRT Model Intuition

Modern approach: use an **item response theory** (IRT) model to jointly estimate aptitude and test question characteristics.

IRT Model Intuition

Modern approach: use an **item response theory** (IRT) model to jointly estimate aptitude and test question characteristics.

Intuition: probability of getting a question right is related to three things

1. the respondent's aptitude
2. the difficulty of the question
3. how well the question measure aptitude

IRT Model Intuition

Modern approach: use an **item response theory** (IRT) model to jointly estimate aptitude and test question characteristics.

Intuition: probability of getting a question right is related to three things

1. the respondent's aptitude
2. the difficulty of the question
3. how well the question measure aptitude

Use a statistical model to estimate all three things at once. Tells us *how much* we should weight each question.

IRT Model Equation

The standard IRT model:

$$p(y_{ij} = 1) = f\left(\beta_j A_i - \alpha_j\right)$$

- A_i is aptitude

IRT Model Equation

The standard IRT model:

$$p(y_{ij} = 1) = f\left(\beta_j A_i - \alpha_j\right)$$

- A_i is aptitude
- α_j is the “difficulty” of question j. High $\alpha_j \rightsquigarrow$ low probability of getting the question right

IRT Model Equation

The standard IRT model:

$$p(y_{ij} = 1) = f\left(\beta_j A_i - \alpha_j\right)$$

- A_i is aptitude
- α_j is the "difficulty" of question j. High $\alpha_j \rightsquigarrow$ low probability of getting the question right
- β_j is the "discrimination" of question j. High $\beta_j \rightsquigarrow$ high-aptitude respondents much more likely to answer correctly than low-aptitude respondents

IRT Model Equation

The standard IRT model:

$$p(y_{ij} = 1) = f\left(\beta_j A_i - \alpha_j\right)$$

- A_i is aptitude
- α_j is the "difficulty" of question j. High $\alpha_j \rightsquigarrow$ low probability of getting the question right
- β_j is the "discrimination" of question j. High $\beta_j \rightsquigarrow$ high-aptitude respondents much more likely to answer correctly than low-aptitude respondents
- f is an increasing function that outputs a value between 0 and 1. E.g., logit or probit function

IRT Model Equation

The standard IRT model:

$$p(y_{ij} = 1) = f\left(\beta_j A_i - \alpha_j\right)$$

- A_i is aptitude
- α_j is the "difficulty" of question j. High $\alpha_j \rightsquigarrow$ low probability of getting the question right
- β_j is the "discrimination" of question j. High $\beta_j \rightsquigarrow$ high-aptitude respondents much more likely to answer correctly than low-aptitude respondents
- f is an increasing function that outputs a value between 0 and 1. E.g., logit or probit function

Many software options to estimate this sort of model; details beyond scope of this class.

Using the Model to Measure Ideology

$p(y_{it} = \text{conservative}) = f\left(\beta_j V_i - \alpha_j\right)$

- "aptitude" $V_i \rightsquigarrow$ ideology (higher values are more conservative)

Using the Model to Measure Ideology

$$p(y_{it} = \text{conservative}) = f\left(\beta_j V_i - \alpha_j\right)$$

- "aptitude" $V_i \rightsquigarrow$ ideology (higher values are more conservative)
- "discrimination" $\beta_j \rightsquigarrow$ how strongly related to ideology is policy question j?

Using the Model to Measure Ideology

$p(y_{it} = \text{conservative}) = f\,(\beta_j V_i - \alpha_j)$

- "aptitude" $V_i \rightsquigarrow$ ideology (higher values are more conservative)
- "discrimination" $\beta_j \rightsquigarrow$ how strongly related to ideology is policy question j?
- "difficulty" $\alpha_j \rightsquigarrow$ how (un)popular is policy question j

Using the Model to Measure Ideology

$p(y_{it} = \text{conservative}) = f\left(\beta_j V_i - \alpha_j\right)$

- “aptitude” $V_i \rightsquigarrow$ ideology (higher values are more conservative)
- “discrimination” $\beta_j \rightsquigarrow$ how strongly related to ideology is policy question j?
- “difficulty” $\alpha_j \rightsquigarrow$ how (un)popular is policy question j

Using the Model to Measure Ideology

$p(y_{it} = \text{conservative}) = f\left(\beta_j V_i - \alpha_j\right)$

- "aptitude" $V_i \rightsquigarrow$ ideology (higher values are more conservative)
- "discrimination" $\beta_j \rightsquigarrow$ how strongly related to ideology is policy question j?
- "difficulty" $\alpha_j \rightsquigarrow$ how (un)popular is policy question j

A potential problem: we don't always know which answer is the "conservative" answer. In notation, we don't know if β_j is positive or negative.

Using the Model to Measure Ideology

$p(y_{it} = \text{conservative}) = f\left(\beta_j V_i - \alpha_j\right)$

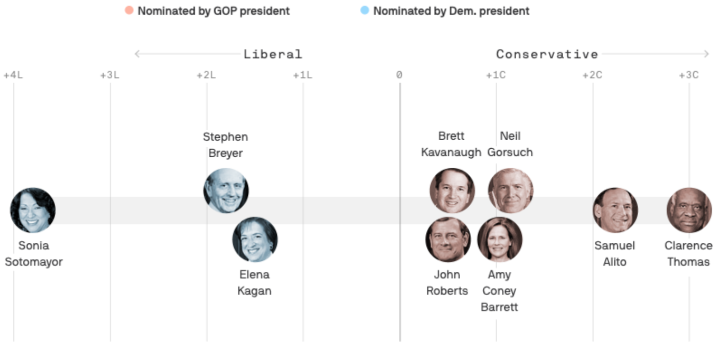
- "aptitude" $V_i \rightsquigarrow$ ideology (higher values are more conservative)
- "discrimination" $\beta_j \rightsquigarrow$ how strongly related to ideology is policy question j?
- "difficulty" $\alpha_j \rightsquigarrow$ how (un)popular is policy question j

A potential problem: we don't always know which answer is the "conservative" answer. In notation, we don't know if β_j is positive or negative.

Solution: it's not actually a problem $\rightsquigarrow$ the model will figure out which items have the same sign of β_j. It's arbitrary whether we say high values of V_i is liberal or conservative, so we're free to pick.

IRT Model

- Actually used by the ETS to score the SAT and other tests
- Widely used by political scientists to estimate ideology and other latent traits
- Better grounded theoretically, but often gives similar answers as simpler approaches
- Useful to examine α_j and β_j to see which questions are useful/which aren't as useful



Ideological scores of Supreme Court justices

Short Authoritarianism Scale

Although there are a number of qualities that people think children should have, every person thinks that some are more important than others. Although you may feel that both qualities are important, please tell me which one of each pair you think is more important for a child to have.

1. *Would you say that it is more important for a child to be INDEPENDENT or RESPECTFUL OF THEIR ELDERS?*
2. *Would you say that it is more important for a child to be OBEDIENT or SELF-RELIANT?*
3. *Would you say that it is more important for a child to be WELL-BEHAVED or CONSIDERATE?*
4. *Would you say that it is more important for a child to be CURIOUS or GOOD MANNERED?*

Additional items in Englehardt et al. (2021):

- Free-spirited vs. Polite
- Orderly vs. Imaginative
- Adaptable vs. Disciplined
- Loyal vs. Open-minded

Four-Item Scale Doesn't Cover Extremes

Discrimination ≈ slope of line

Difficulty ≈ value of authoritarianism scale where $p(y_{ij} = \text{authoritarian}) = 0.5$

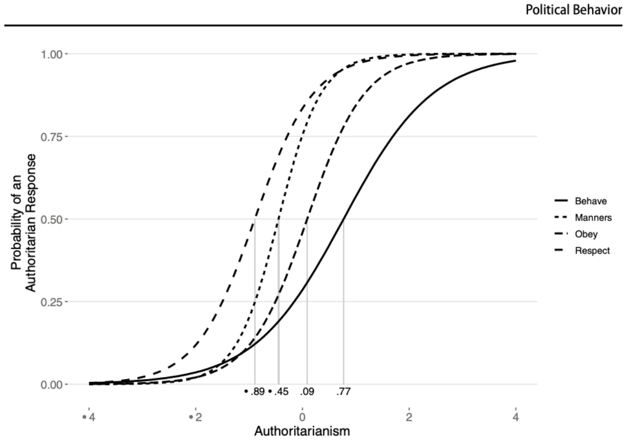


Fig. 1 Item response curves for four child values items, 2016 American National Elections Study Data

Additional Items Cover More of Scale

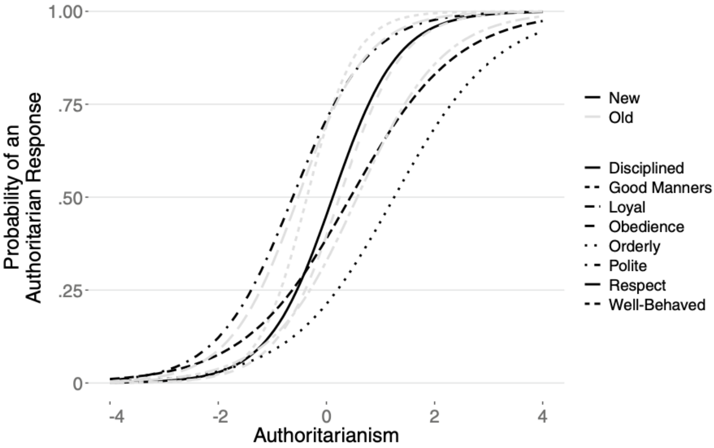


Fig. 2 Item response curves, Qualtrics study

Summing Up

- Asking several questions can help get a better measure of latent traits
- Several ways to generate a scale: averaging items, averaging z-scores, IRT model
- Even more are out there: factor models, principal components analysis, inverse covariance weighting, ...