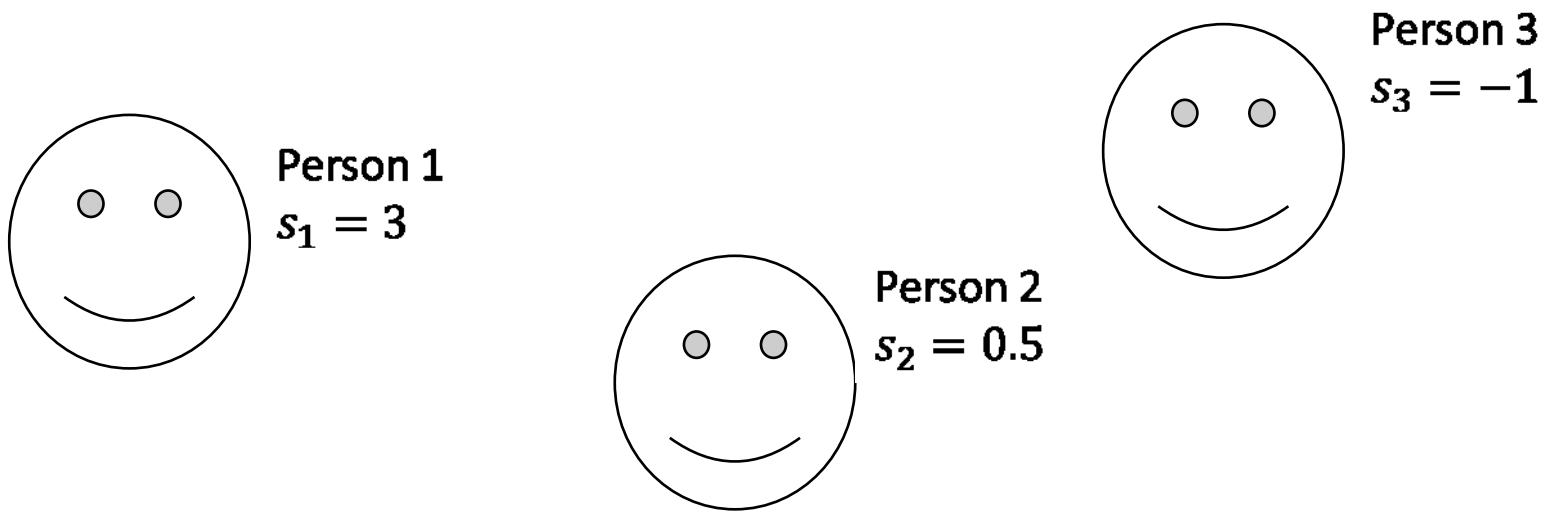
Conceptual overview

PSYn5440 – Introduction to Factor Analysis

Week 1

Basic principles

- Let's begin with a little thought experiment
- Imagine you gather a bunch of people in a room and let everyone secretly draw a number from a hat
- Let's call each person *i*'s drawn number their *secret score*, s_i



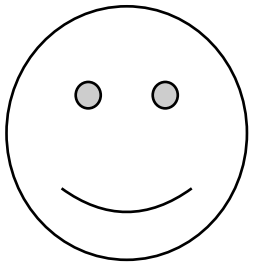
Basic principles

- **Now, in this room there are three booths**
- **Each booth contains instructions which are not known to you**
- **You ask each person to visit all three booths and follow the instructions**

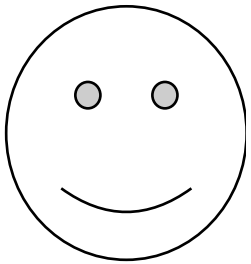
- **The first booth contains the following instructions:**

Take your secret score, multiply it by 0.3, add 1, and say the result out loud

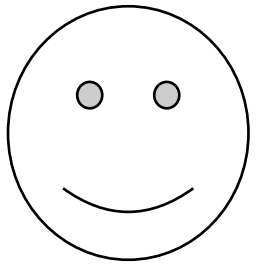
(in other words, $x_{i1} = 1 + 0.3 * s_i$)



Person 1
$s_1 = 3$
"1.9"



Person 2
$s_2 = 0.5$
"1.15"



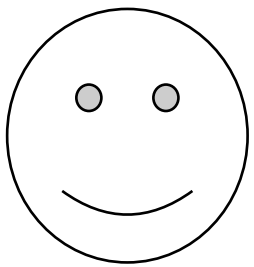
Person 3
$s_3 = -1$
"0.7"

Basic principles

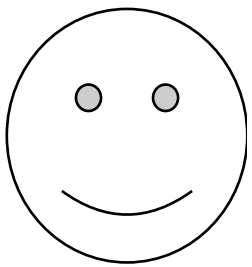
- **The second booth contains the following instructions:**

Take your secret score, multiply it by 1.2, add 8, and say the result out loud

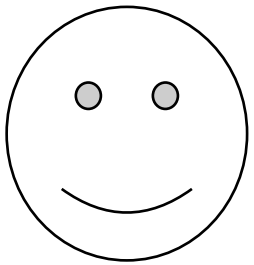
(in other words, $x_{i2} = 8 + 1.2 * s_i$)



Person 1
$s_1 = 3$
"11.6"



Person 2
$s_2 = 0.5$
"8.6"



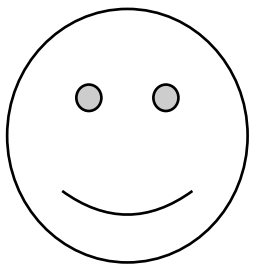
Person 3
$s_3 = -1$
"6.8"

Basic principles

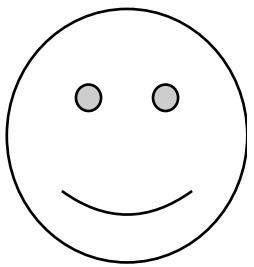
- **The third booth contains the following instructions:**

Take your secret score, multiply it by -0.8, subtract 3, and say the result out loud

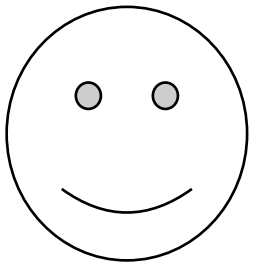
(in other words, $x_{i3} = -3 - 0.8 * s_i$)



Person 1
$s_1 = 3$
"-5.4"



Person 2
$s_2 = 0.5$
"-3.4"



Person 3
$s_3 = -1$
"-2.2"

Basic principles

- **As you can see, the numbers reported to us were always *functions* of each person's *secret score*, s_i**

- **Suppose we write this all down into a little data matrix**

Person	Booth 1	Booth 2	Booth 3
1	1.9	11.6	-5.4
2	1.15	8.6	-3.4
3	0.7	6.8	-2.2

Basic principles

- This is an example of **multivariate data** – data for a sample of individuals on a number of variables
- The data matrix contains scores on three variables we were able to directly *observe* (well, each person told us truthfully). In the world of factor analysis, we refer to these observable or measurable variables as **manifest variables** (MVs)

One column for each variable

One row for each person

Person	Booth 1	Booth 2	Booth 3
1	1.9	11.6	-5.4
2	1.15	8.6	-3.4
3	0.7	6.8	-2.2

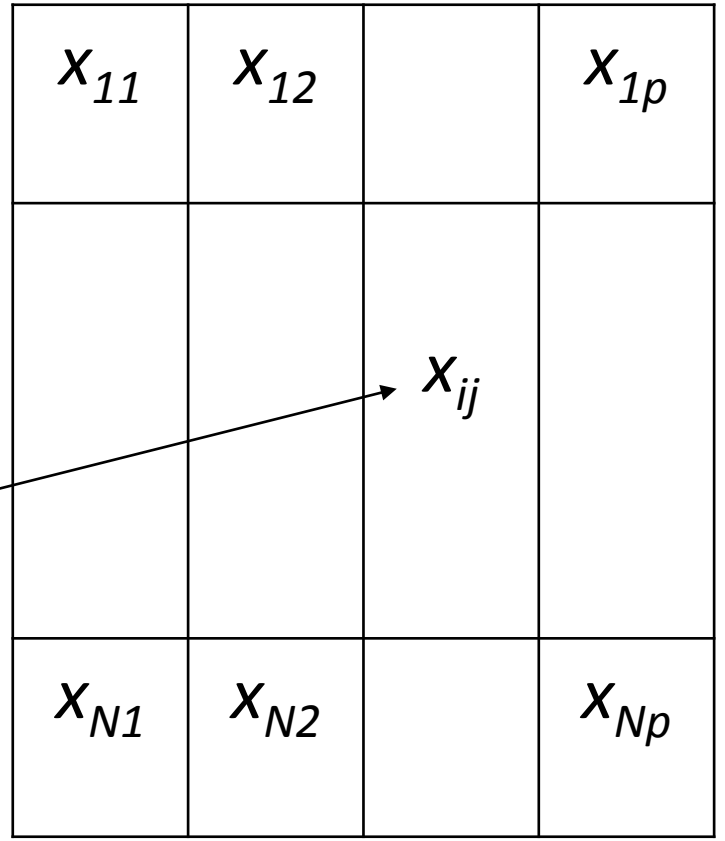
Basic principles

Data matrix:

p columns (variables)

X **=** *N rows (individuals)*

Score of person *i* on variable *j*



Basic principles

- **What can we observe in these data?**
- Well, we can try to compare the persons
  - Person 1 always seems to have the most extreme scores
  - Person 3 *kinda* looks like they always have half the scores Person 1 has?

Person	Booth 1	Booth 2	Booth 3
1	1.9	11.6	-5.4
2	1.15	8.6	-3.4
3	0.7	6.8	-2.2

Basic principles

- **What can we observe in these data?**
- We can also try to look at each variable

Person	Booth 1	Booth 2	Booth 3
1	1.9	11.6	-5.4
2	1.15	8.6	-3.4
3	0.7	6.8	-2.2

  - Booth 2 seems to have the largest absolute values
  - Booth 3 is always negative
  - The sample range of Booth 1 is 1.2
  - Maybe we could look at the variation of each variable over the sample? Like a standard deviation?

Basic principles

- **What can we observe in these data?**
- We can also look at pairs of variables!
  - How do the variables correlate / covary across our sample?

	Booth 1	Booth 2	Booth 3
Booth 1	1	r_{12}	r_{13}
Booth 2	r_{21}	1	r_{23}
Booth 3	r_{31}	r_{32}	1

- If we calculate correlations for each variable pair, we can arrange the correlation values into a **correlation matrix**

Basic principles

Correlation matrix:

R:

p manifest variables

1	r_{12}	r_{13}			r_{1p}
r_{21}	1	r_{23}			r_{2p}
r_{32}	r_{32}	1			r_{3p}
				r_{kj}	
			r_{jk}		
r_{p1}	r_{p2}	r_{p3}			1

p manifest variables

Basic principles

- Now, it doesn't make a lot of sense to calculate correlation of two variables with three observations each…
- But if our thought experiment contained more people in our sample, we would be able to observe that the manifest variables do, in fact, **correlate** / **covary**

- But why might that be? Why do the numbers or scores from the three booths correlate?

Basic principles

- One reason for the observed correlations might be that the three manifest variables have a **common cause**

- In other words, maybe they are all caused of affected by the **same thing**?

- In fact, we know this is true. The variables were all functions of each person's *secret score*

Basic principles

- The *secret score* in our little thought experiment was something that is, in the world of factor analysis, referred to as a **latent variable**

- Latent variables are “hidden” and unobserved (nobody told us their secret score), yet they affect the **manifest variables** in some systematic way

Basic principles

- We might not be able to know what each person's *secret score* was…
- …but maybe we can try to indirectly infer on it if we study the relationships between the variables we *have* observed (i.e., the **manifest variables**)

Basic principles

- This is the fundamental idea behind **factor analysis**
- Observed (manifest) variables that correlate do so because they share a common cause, some unseen, hidden, or latent variable

- In factor analysis, latent variables are also called **factors**

Key terms and definitions

- **Manifest variable** – variable that can be directly measured (or observed)
- **Latent variable** – variable that cannot be directly measured (or observed) – a hypothetical construct. A latent variable is a **factor** in factor analysis. Thus, a factor is a variable and individuals have scores on those factors (hypothetically).
- **Population** – The entire set of individuals of interest
- **Sample** – A selected group of individuals from the population (N persons)

Basic principles – cont'd

- In our thought experiment, we had everyone draw a *secret score*
- Therefore, we could easily hypothesize that the answers from the booths varied because they had something to do with the (also varying) *secret score*

- However, that was just a (silly) thought experiment. The world is, unfortunately, much more complicated and we are much less informed

Basic principles – cont'd

- The fundamental principle still stands, though:

Observed (manifest) variables correlate **because** they are **affected** by the same unobserved (latent) variables, or **factors**.

In other words, the **structure of the correlation** (or covariance) **matrix** can be described or explained by the existence of latent variables.

- Of course, this reasoning cannot be applied to just *any* correlation matrix. The hypothetical factor(s) must be theoretically justifiable.

Basic principles – cont'd

- The **objective** of factor analysis, then, is to **uncover** and **understand** the structure that produces the correlations in the data

- **Assumption** - there exists a *small number* of factors (within a particular domain) which influence the MVs and thus produce the correlations (covariances) between manifest variables. If this were not the case, we would gain very little by doing factor analysis.

- e.g., it is assumed that a **limited** number of mental abilities will explain relationships between **all** ability tests. No MV single-handedly represents a distinct ability or trait.

Basic principles – cont'd

- As said before, we assume that the factors **influence or affect** the MVs.
- Our aim, then, is not only to uncover *how many* factors cause the observable correlation in our data, but also *how each factor affects* each manifest variable (in fact, these are two questions you can hardly separate)
- The degree of a factor's influence is represented by the so-called **factor loading**

Basic principles – cont'd

- **Let's briefly revisit the booth instructions from our thought experiment:**

1) ***Take your secret score, multiply it by 0.3, add 1, and say the result out loud***
   (in other words, $x_{i1} = 1 + 0.3 * s_i$)
2) ***Take your secret score, multiply it by 1.2, add 8, and say the result out loud***
   (in other words, $x_{i2} = 8 + 1.2 * s_i$)
3) ***Take your secret score, multiply it by -0.8, subtract 3, and say the result out loud***
   (in other words, $x_{i3} = -3 - 0.8 * s_i$)

Basic principles – cont'd

- **These would be akin to the factor loadings**

1) *Take your secret score, multiply it by 0.3, add 1, and say the result out loud*

(in other words, $x_{i1} = 1 + 0.3 * s_i$)

2) *Take your secret score, multiply it by 1.2, add 8, and say the result out loud*

(in other words, $x_{i2} = 8 + 1.2 * s_i$)

3) *Take your secret score, multiply it by -0.8, subtract 3, and say the result out loud*

(in other words, $x_{i3} = -3 - 0.8 * s_i$)

Basic principles – cont'd

- The numerical values of factor loadings indicate the **strength** of the factor's influence on the MV (a zero indicates no influence). Factor loadings are equivalent to **regression coefficients**, standing for the influence of a factor (independent variable) on a MV (dependent variable)

- Let's say that we have 'discovered' a factor which causes our MVs to correlate. We have somehow obtained the factor loadings. Now, how do we decide what *is* this factor? What theoretical idea or construct does this factor / latent variable represent?

- The pattern of factor loadings helps us determine the nature of a factor

…in other words, a factor is defined by the subset of MVs that it substantially influences

Example

- Suppose we have scores from a sample of individuals on 4 performance measures: paragraph comprehension, vocabulary, arithmetic skills, and mathematical problem solving. We get the following correlation matrix:

	PC	VO	AR	MPS
PC	1			
VO	.49	1		
AR	.14	.07	1	
MPS	.48	.42	.48	1

Example

- We would like to identify the underlying factors to explain the correlations. Thus, we employ factor analysis methods and obtain a factor loading matrix:

	Factor 1	Factor 2
PC	.70	.10
VO	.70	.00
AR	.10	.70
MPS	.60	.60

Example

	Factor 1	Factor 2
PC	.70	.10
VO	.70	.00
AR	.10	.70
MPS	.60	.60

- Elements in the matrix represent the linear influence of each factor on each measure.

- In this course, we will study methods that will allow us to obtain such interpretable factor loading matrices.

- Keep in mind that we are using a model – a one which represents some hypothesized structure of observed data. Any mathematical model is – at least to some extent – wrong and does not perfectly correspond to reality.
- A model that makes sense conceptually but does not fit reasonably well is useless.
- A model that fits great but does not make sense is useless as well.

- A factor analysis is not applicable to just any data.

- In the world of factor analysis, situations differ regarding the existence of prior hypotheses / knowledge about the number and nature of the factors:

Exploratory (unrestricted) FA:

We have little prior idea of how many and what kind of factors there are.

Confirmatory (restricted) FA:

We do have a hypothesis (or hypotheses) about the number and nature of factors.

...the underlying theoretical model is **the same**!

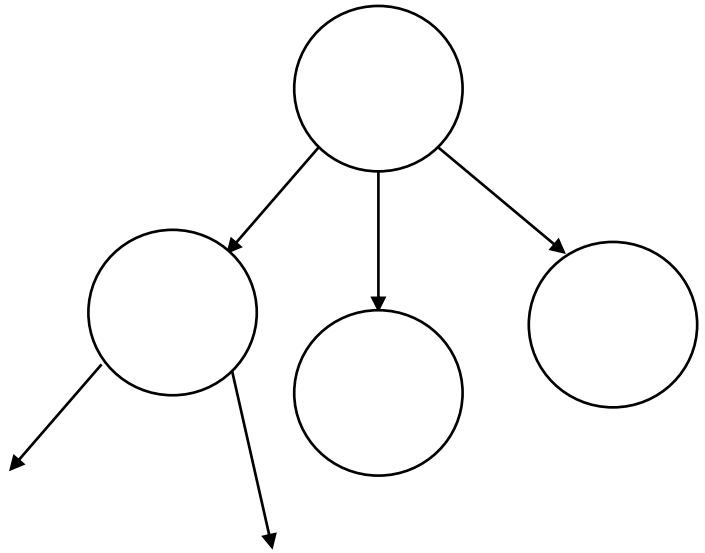
A bit of history

- Factor analysis began with the study of mental abilities

- Charles Spearman proposed the first factor model in 1904:
  - Performance on any test is a function of two factors – a **general** ability factor (Spearman's **g**) common to all ability tests, and a **specific** ability factor relevant only to the specific test in question.
  - → The two-factor theory of intelligence
  - Ability tests correlate because they all depend on the general factor.

A bit of history

- Burt and Vernon, on the other hand, proposed a **hierarchical model** of human abilities:
  - The human mind is organized in a hierarchy of abilities.
  - The general ability sits atop this hierarchy
  - More specific abilities are located lower in the hierarchy



A bit of history

- The **Common Factor Model** of L. L. Thurstone became the most prominent approach to FA since the 1940s. Thurstone disagreed with both the notion of **g** and a hierarchy of abilities.
- According to Thurstone, MVs depend on two kinds of underlying factors:
  - **Common factors** that are *common* to more than one MV
  - **Unique factors** that influence only one MV. Unique factors do not explain correlations between MVs.
- The p manifest variables depend on m common factors and p unique factors, where $m < p$

The Common Factor Model

- **How can we understand this common + unique factor business?**
- **Let's get back to our thought experiment and tweak it a bit**
- **Let's say that each booth *j* would contain two dice and the following generalized instructions:**

Take your secret score, multiply it by g, and add f. Then, roll the two dice, sum up the pips to get h, and add this to the result.

(in other words, $x_{ij} = f_j + g_j * s_i + h_{ij}$)

The Common Factor Model

- This dice roll would be analogous to the **unique factor** (well not *really*, but bear with me for now)
- As you can see, this value doesn't depend on the secret score, but it's still plausibly different for every participant *i*
- The dice roll is also different for every booth *j*, so dice in one booth do not affect the result in any other booth

Take your secret score, multiply it by g, and add f. Then, roll the two dice, sum up the pips to get h, and add this to the result.

(in other words, $x_{ij} = f_j + g_j * s_i + h_{ij}$)

The Common Factor Model

- Dice rolls are purely random, so they represent **random error**
- However, the **unique factors** are a bit more complicated – they do not represent only (unsystematic) random error
- They also represent systematic effects that affect only the one particular MV in question – this is called the **specific factor**
- Because the specific factor affects only one MV, we cannot disentangle it from the random error
- **Bottom line** - each unique factor has two components:
  - Specific factor
  - Error of measurement

The Common Factor Model

We can break down the variance of a given MV in the following way:

Observed variance = Common variance + Unique variance

…and because:

Unique variance = Specific variance + Error variance

…then:

Observed variance = Common variance + Specific variance + Error variance

The Common Factor Model

Observed variance = Common variance + Specific variance + Error variance

$$\textbf{Communality} = \frac{\text{Common variance}}{\text{Observed variance}} = 1 - \frac{\text{Unique variance}}{\text{Observed variance}}$$

= the proportion of observed variance due to common factors

The Common Factor Model

- Let's (for the final time, I guess, sniff) revisit the thought experiment with the dice plot twist one more time:

Take your secret score, multiply it by g, and add f. Then, roll the two dice, sum up the pips to get h, and add this to the result.

$$(x_{ij} = f_j + g_j * s_i + h_{ij})$$

x_{ij} - the number person *i* shouts out of booth *j*

f_j - some constant for booth *j*

g_j - multiplier of the secret score s_i for booth *j*

h_{ij} - result of random dice roll of person *i* in booth *j*

The Common Factor Model

The mathematical expression of the Common Factor Model:

$$x_{ij} = \mu_j + \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \cdots + \lambda_{jm}z_{im} + 1u_{ij}$$

Mean + Common factor part + Unique factor part

x_{ij} **is the score of person *i* on manifest variable *j***

μ_j **is the mean of manifest variable *j***

The Common Factor Model

The mathematical expression of the Common Factor Model:

$$x_{ij} = \mu_j + \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \cdots + \lambda_{jm}z_{im} + 1u_{ij}$$

Mean + Common factor part + Unique factor part

z_{ik} **is the common factor score of person *i* on factor *k* (latent variable score)**

λ_{jk} **is the factor loading (regression weight) of MV *j* on factor *k***

u_{ij} **is the unique factor score of person *i* on unique factor *j* (also latent)**

...the unique factor score consists of a specific part and an error part:

$$u_{ij} = s_{ij} + e_{ij}$$

The Common Factor Model

We mentioned variances before – do not confuse the scores (x_{ij}, z_{ik}...) with the variances of those scores [var(x_j), var(z_k)], which is how these scores vary across persons.

The model can be re-written by subtracting the mean from both sides:

$$x_{ij} - \mu_j = \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \cdots + \lambda_{jm}z_{im} + 1u_{ij}$$

...thus, we can see that the model specifies the deviation from the mean as a function of the common and unique factors.

The Common Factor Model

Important assumption: In the model, the unique factor scores for different MVs are assumed to be uncorrelated over all persons. Therefore, all partial correlations between MVs, controlling for the effect of the common factors, are assumed to be zero.

- In other words, correlations between MVs are only due to the common factors (that's why they're called *common*)
- This assumption refers to the population

The Common Factor Model

- What factors are common and what factors are specific depends on the manifest variables in the dataset.
- If we change the set of MVs by introducing new MVs or deleting MVs, we can potentially change specific factors into common factors, and so on.

The Common Factor Model

- The model is will always be wrong to some degree (it's a *model* after all). What are some of the ways the model could be wrong?
  - 1) The assumption of linearity – the MVs are specified as linear functions of factors. Nobody really thinks the real world is perfectly linear.
  - 2) The number of common factors is generally assumed to be small ($m << p$). In reality, there are probably many, many common influences on a score. However, we hope to identify the non-negligible ones.
- We should recognize the common factors will not perfectly explain the variation and covariation of the manifest variables.

The Common Factor Model

- The model equation looks like a multiple regression equation.
  - The manifest variables are dependent variables
  - The factors are independent variables
  - The factor loadings are regression weights / coefficients

- The factor analysis model is like a set of multiple linear regressions where the independent variables are unobservable.