Exploratory Factor Analysis
（探索的因子分析）

Yasuyo Sawaki
Waseda University
JLTA2011 Workshop
Momoyama Gakuin University
October 28, 2011

Today's schedule

Part 1: EFA basics

- Introduction to factor analysis
- Logic behind EFA
- Key steps of EFA
- EFA vs. CFA

(BREAK)

Part 2: EFA Exercise

What is factor analysis?

- A multivariate statistical technique developed in the early 1900s
- A method to examine relationships between observed variables（観測変数）and latent factors（潜在因子）

Types of factor analysis

- Exploratory factor analysis (EFA; 探索的因子分析)
  - Data-driven approach
  - Often used in early stages of an investigation
- Confirmatory factor analysis (CFA; 検証的/確認的因子分析)
  - Theory-driven approach
  - Often used in later stages of an investigation to confirm specific hypotheses

- Combining EFA and CFA

Previous applications of EFA in applied linguistics (1)

- Development and validation of survey and assessment instruments

Example 1: Bachman, Davidson, Ryan, & Choi (1995)

- Examining comparability of the factor structures of two English language tests (Cambridge FCE and TOEFL)
- EFAs for each test separately, followed by another EFA run for a combined analysis of the two tests

Previous applications of EFA in applied linguistics (2)

- Development and validation of survey and assessment instruments

Example 2: Vandergrift, Goh, Mareschal, & Tafagodtari (2006)

- Developing and validating a new survey on L2 learners' metacognitive awareness and strategy use in listening comprehension
- An initial EFA to finalize survey content, followed by a CFA on a different sample

Previous applications of EFA in applied linguistics (3)

- Descriptive use of EFA

Example: Biber, Conrad, Reppen, Byrd, & Helt (2002)

- An EFA of an academic English language corpus to identify linguistic characteristics of various spoken and written registers

Logic behind factor analysis

- Both EFA and CFA based on the **common factor model（共通因子モデル）**

Underlying logic: Variables correlate because they tap into the same construct(s) to certain degrees

Goal: To identify an optimal number of latent factors (factor solution) that describes the pattern of relationships among a set of variables sufficiently well (reproduction of the observed correlation matrix)

The common factor model

- Decomposing variance of observed variables into two parts:
  - *Common variance:* part of variance influenced by a latent factor shared across different variables (common factors; 共通因子)
  - *Unique variance:* part of variance not explained by the common factors
    - Variance explained by factors other than the common factors (unique factors; 独自因子）
    - Variance due to measurement error

EFA: Issues

- Readily available in statistical packages
- Easy to implement, but careful consideration of various issues in different steps of the analysis is required to obtain interpretable and meaningful analysis results

Key steps of EFA

Step 1: Checking appropriateness of study design and data type for conducting EFA

Step 2: Deciding on the number of factors to extract

Step 3: Extracting and rotating factors

Step 4: Interpreting key EFA results

A sample scenario

- An EFL speaking and listening test
- Sample size: N=200
- Variables 1-6 for speaking, and Variables 7-12 for listening (4-6 score points available for each variable)
- Expected findings
  - There are two latent factors: One each for speaking and listening
  - The two factors are correlated with each other because they are both different aspects of L2 language ability
- Software: SPSS (PASW Statistics Version 18)

Step 1: Checking appropriateness of study design and data type for EFA (1)

Data type: determines the type of correlation matrix to be analyzed

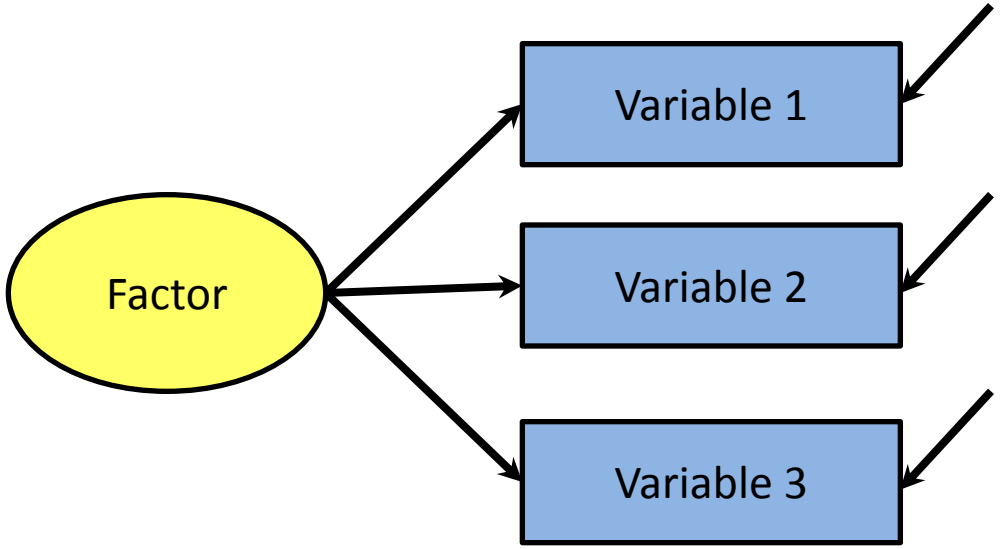
Common examples

- Interval or quasi-interval scales → Pearson correlation matrix
- Ordered categorical data
  - Dichotomous data → tetrachoric correlation matrix
  - Polytomous data → polychoric correlation matrix

Step 1: Checking appropriateness of study design and data type for EFA (2)

Number of variables: At least 3 variables needed per factor

Example 1: 3 variables to identify one factor
Number of data points available in a correlation matrix with k=3:
k(k+1)/2 = 3(3+1)/2 = 6



Just identified model

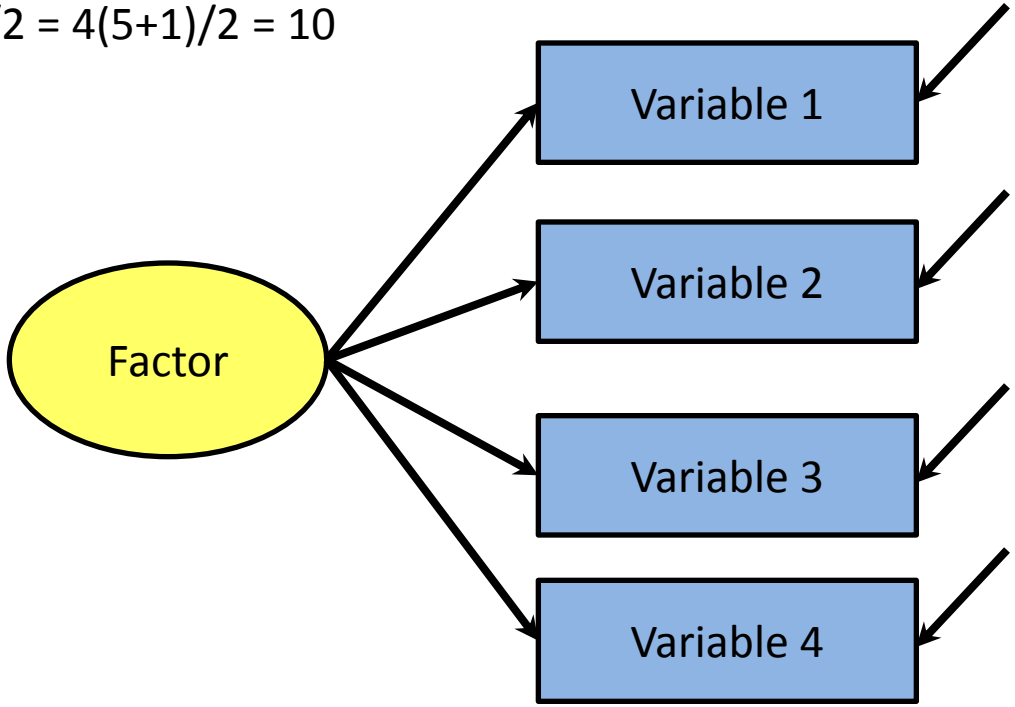
Step 1: Checking appropriateness of study design and data type for EFA (3)

Number of variables: At least 3 variables needed per factor

Example 2: 4 variables to identify one factor

Number of data points available in a correlation matrix with k=4 :

k(k+1)/2 = 4(5+1)/2 = 10

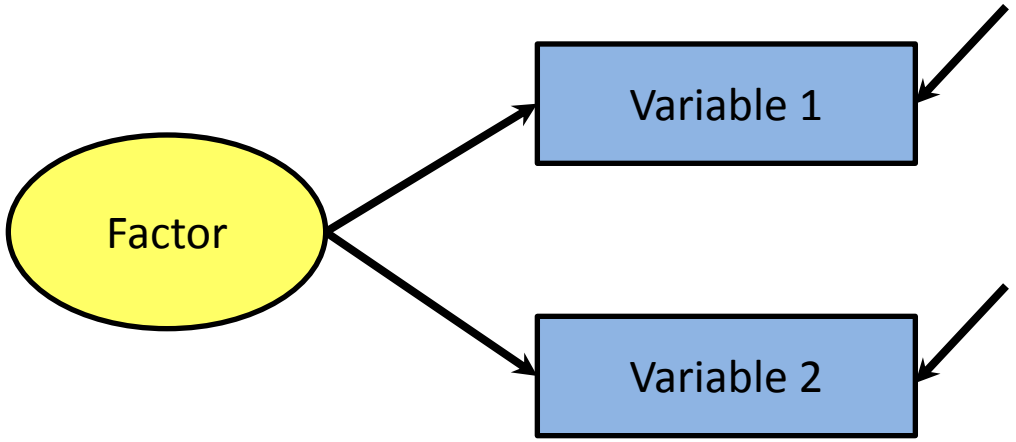


Over identified model

Step 1: Checking appropriateness of study design and data type for EFA (4)

Number of variables: At least 3 variables needed per factor

Example 2: 4 variables to identify one factor
Number of data points available in a correlation matrix with k=2 :
k(k+1)/2 = 2(3+1)/2 = 3



Under identified model

Step 1: Checking appropriateness of study design and data type for EFA (5)

Sample size: Suggestions about sample size in the literature vary

A required sample size depends on many factors such as:

- Strength of correlations between factors and their indicator variables
- Reliability
- Score distribution of variables (e.g., normality)
- Number of factors to extract

SEE: Fabriger et al. (1999), Floyd & Widaman (1995), Tabachnick & Fidell (2007)

Step 2: Deciding on the number of factors to extract(1)

Various approaches to determining the number of factors

- Approaches based on eigenvalues for the correlation matrix

  Eigenvalue（固有値）: Shows how much of the variance of a set of variables can be explained by a factor

- Goodness of model fit（**モデルの適合度**）： goodness-of-fit statistics available for certain estimation methods (e.g., ML)

Step 2: Deciding on the number of factors to extract(2)

- Approaches based on eigenvalues for the correlation matrix

  Goal: To identify the number of factors with large enough eigenvalues to explain relationships among observed variables

- Kaiser's criterion (Kaiser, 1960)
- Scree test (Cattell, 1966)
- Parallel analysis (PA; Horn, 1965)

Step 2: Deciding on the number of factors to extract(3)

- Kaiser's criterion (Kaiser, 1960)
  - The number of factors to extract = the number of factors with eigenvalues exceeding 1.0

説明された分散の合計

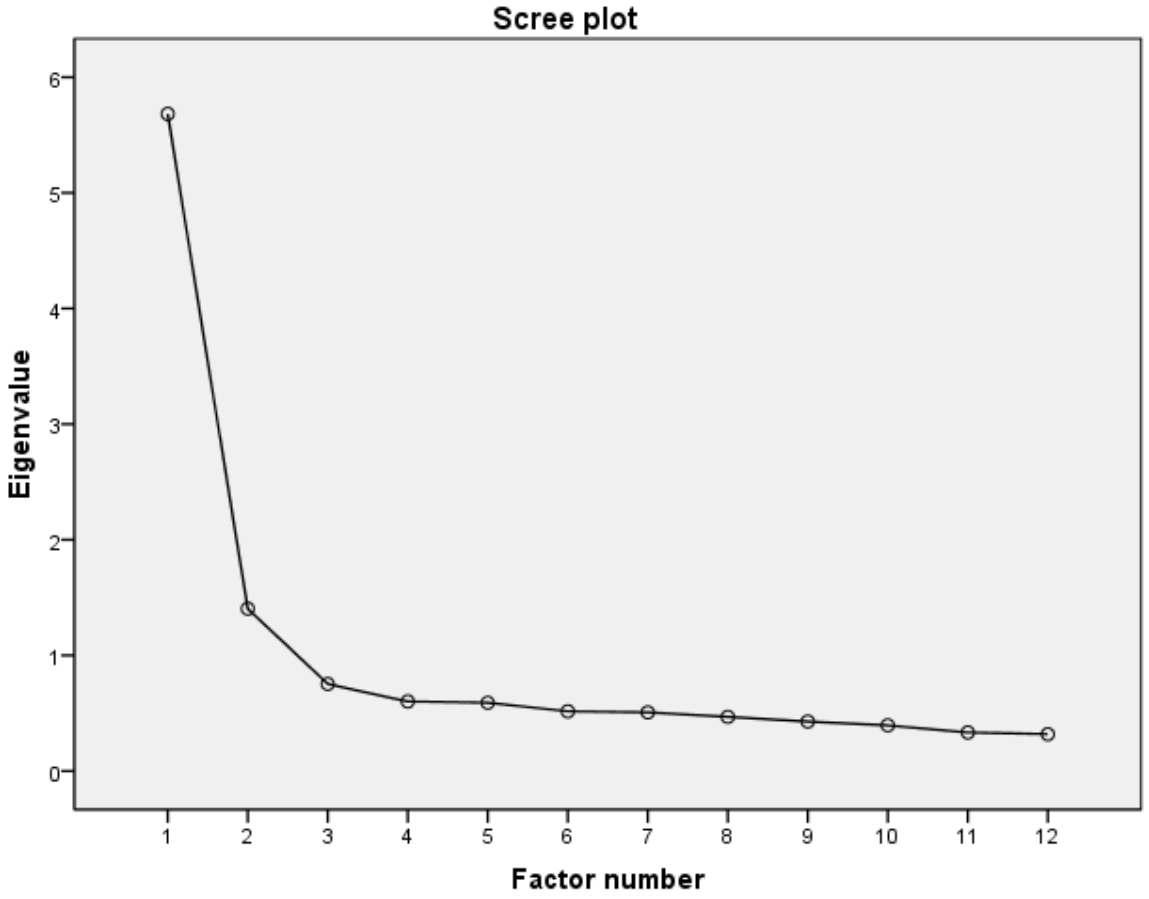
因子	初期の固有値			抽出後の負荷量平方和			回転後の負荷量平方和[a]
	合計	分散の %	累積 %	合計	分散の %	累積 %	合計
1	5.683	47.358	47.358	5.209	43.412	43.412	4.625
2	1.404	11.697	59.054	.926	7.715	51.127	4.339
3	.753	6.277	65.331				
4	.602	5.014	70.346				
5	.591	4.921	75.267				
6	.517	4.305	79.572				
7	.507	4.228	83.800				
8	.468	3.902	87.702				
9	.429	3.571	91.273				
10	.395	3.291	94.564				
11	.333	2.779	97.343				
12	.319	2.657	100.000				

因子抽出法: 主因子法

a. 因子が相関する場合は、負荷量平方和を加算しても総分散を得ることはできません。

Step 2: Deciding on the number of factors to extract(4)

- Scree plot (Cattell, 1966)
  - The number of factors to extract = the “elbow” in the graph



Step 2: Deciding on the number of factors to extract(5)

- Parallel analysis (PA; Horn, 1965)
  - Comparison of eigenvalues obtained from real data against those obtained from multiple samples of random numbers (see Hayton,et al., 2004; Liu& Rijmen, 2008 for sample SPSS and SAS programs)

  Steps:

  1. Obtain a scree plot for the actual data being analyzed
  2. Run a program for PA, which calculates eigenvalues for multiple samples of random numbers; then take the mean of eigenvalues for each component across the PA runs
  3. Plot the mean eigenvalues for individual components from the PA over the scree plot for comparing the eigenvalues from the actual data and those from the PA runs

Step 2: Deciding on the number of factors to extract(6)

- Goodness-of-fit statistics
  - Available for certain estimation methods of model parameter estimation

Example: maximum likelihood (ML; 最尤法)

Step 2: Deciding on the number of factors to extract(7)

Example: maximum likelihood (ML; 最尤法)

因子行列[a]

	因子	
	1	2
Speak1	.644	-.115
Speak2	.712	-.198
Speak3	.731	-.183
Speak4	.737	-.284
Speak5	.714	-.344
Speak6	.724	-.170
Listen1	.607	.288
Listen2	.680	.283
Listen3	.451	.253
Listen4	.645	.309
Listen5	.597	.367
Listen6	.598	.408

因子抽出法: 最尤法

a. 2 個の因子が抽出されました。4 回の反復が必要です。

適合度検定

カイ2乗	自由度	有意確率
34.849	43	.807

Step 2: Deciding on the number of factors to extract(8)

- Issues of consideration
  - Some widely used methods are easy to implement (e.g., Kaiser's criterion, scree test)
  - However, different methods have different disadvantages
    - Kaiser's criterion: Often found to produce misleading results
    - Scree plots: Can be difficult to decide where the "elbow" is
    - ML estimation: Data have to satisfy distributional assumptions (deviation from normality → misleading analysis results)

Step 2: Deciding on the number of factors to extract(9)

- Issues of consideration (continued)
  - What matrix should be analyzed when determining the number of factors? (Brown, 2006; Fabriger et al., 1999)
    - Kaiser's criterion: observed correlation matrix required
    - Scree test: possible both on observed and reduced correlation matrices
- What should we do in practice?
  - Use multiple methods
  - In later steps of the analysis, carefully examine substantive interpretability and parsimony of a given factor solution

Step 3: Extracting and rotating factors(1)

Choice of a factor extraction method depends on multiple factors such as data type, distribution, and information you need

Common methods used with continuous variables:

Method	Distributional assumption	Goodness-of-fit statistics
Principal factor analysis	No assumption imposed	Not available
Maximum likelihood (ML)	Normality assumed	Available

Step 3: Extracting and rotating factors(2)

- Principal component analysis (PCA; 主成分分析)
  - NOT based on the common factor model; NOT consistent with the purpose of examining underlying factor structure
  - A mathematical data reduction method; NOT based on the common factor model
  - More appropriate when, for example, the goal is to create a composite variable out of a larger number of factors

Step 3: Extracting and rotating factors(3)

Factor rotation

- **One factor solution:** Results from the initial factor solution can be interpreted
- **A solution with multiple factors:** Results from the initial factor solution is difficult to interpret → **Factor rotation** (a mathematical transformation) is often conducted
  - Orthogonal rotation（直行回転）: Correlations among factors NOT allowed (e.g., Varimax rotation)
  - Oblique rotation（斜交回転）: Correlations among factors allowed (e.g., Oblimin rotation; Promax rotation)

Step 4: Extracting and rotating factors(4)

Factor loadings without rotation (based on PFA)

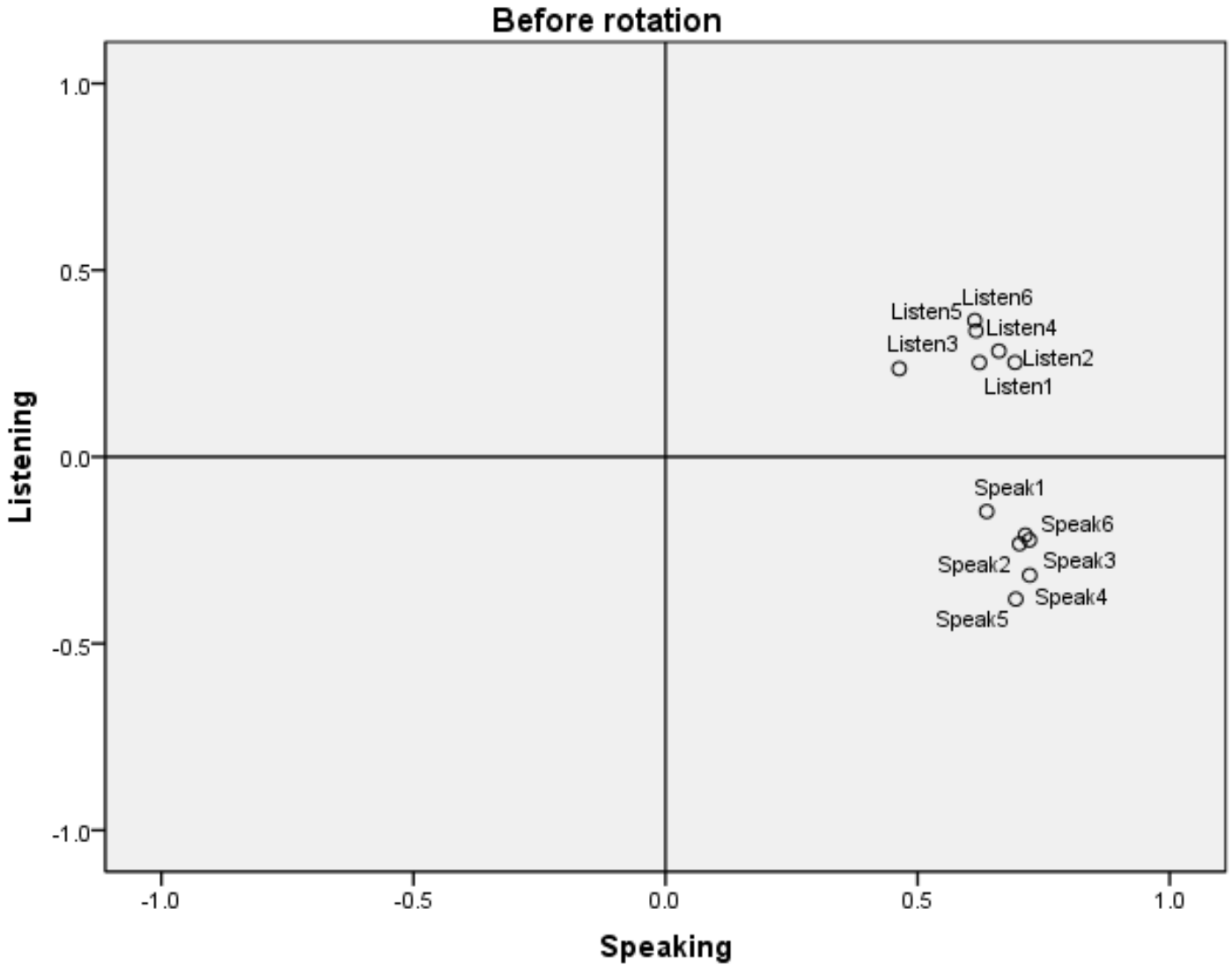
因子行列[a]

	因子	
	1	2
Speak1	.637	-.146
Speak2	.702	-.233
Speak3	.722	-.223
Speak4	.722	-.317
Speak5	.695	-.381
Speak6	.714	-.210
Listen1	.623	.252
Listen2	.693	.253
Listen3	.464	.236
Listen4	.661	.283
Listen5	.616	.337
Listen6	.613	.365

因子抽出法: 主因子法

a. 2 個の因子が抽出されました。6 回の反復が必要です。

Step 4: Extracting and rotating factors(5)



Step 4: Extracting and rotating factors(6)

パターン行列[a]

	因子	
	1	2
Speak1	.558	.133
Speak2	.700	.057
Speak3	.700	.080
Speak4	.813	-.037
Speak5	.872	-.130
Speak6	.680	.092
Listen1	.074	.620
Listen2	.116	.656
Listen3	-.002	.522
Listen4	.060	.678
Listen5	-.031	.723
Listen6	-.066	.756

因子抽出法: 主因子法
回転法: Kaiserの正規化を伴うﾌﾟﾛﾏｯｸｽ法

a. 3回の反復で回転が収束しました。

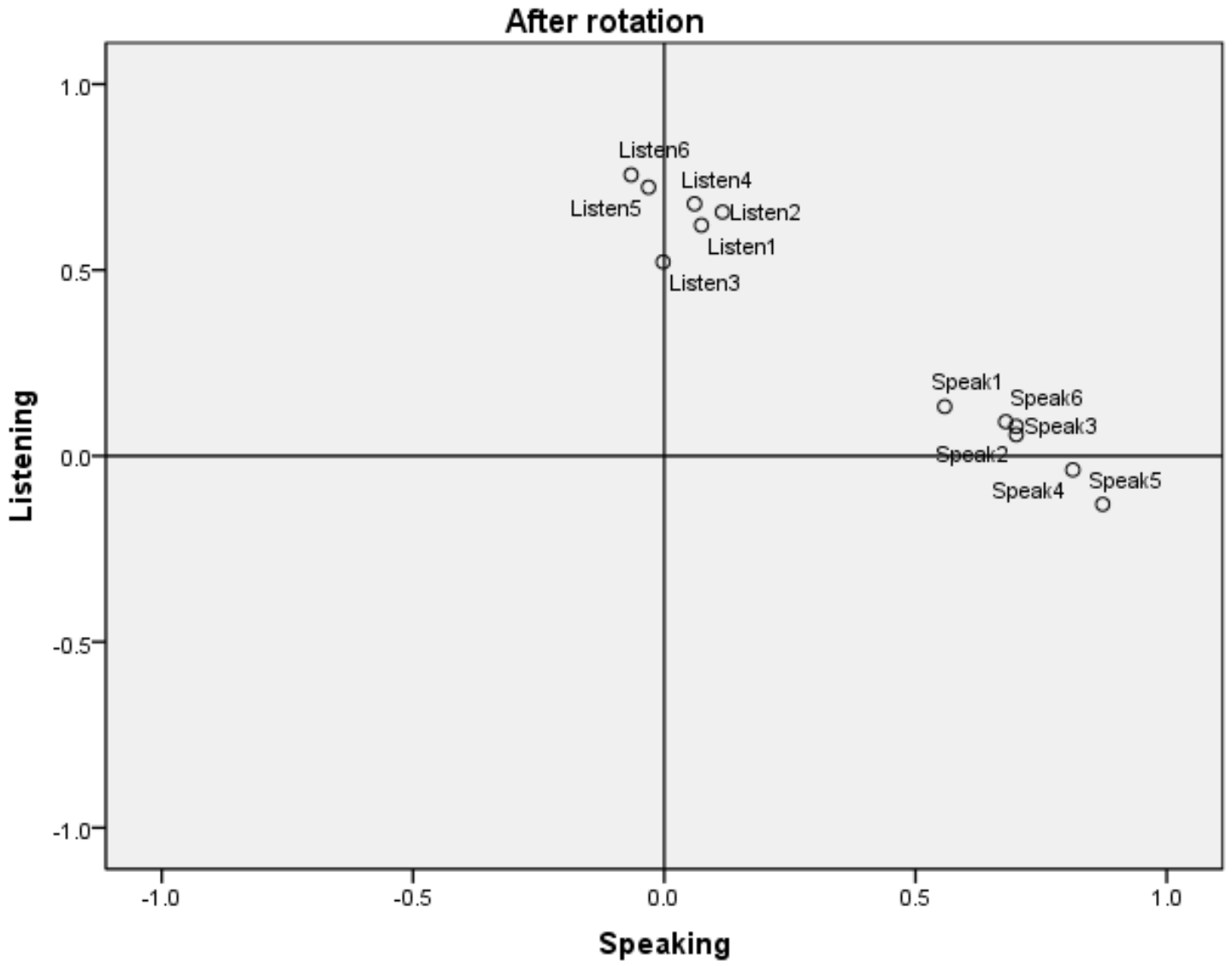
Factor loadings and inter-factor correlation after Promax rotation (based on PFA)

因子相関行列

因子	1	2
1	1.000	.663
2	.663	1.000

因子抽出法: 主因子法
回転法: Kaiserの正規化を伴うﾌﾟﾛﾏｯｸｽ法

Step 4: Extracting and rotating factors(7)



Step 3: Extracting and rotating factors(8)

Issues of consideration

- With language data, which type of factor rotation (orthogonal vs. oblique) is generally preferred?
- Should factors be correlated to use an oblique rotation?

Step 4: Interpreting key analysis results(1)

Key analysis results to focus on:

- **Communality （共通性）:** The estimated proportion of the variance of a given variable explained by the factors included in the model
- **Factor loading（因子負荷量）:** A standardized estimate of the strength of the relationship between an observed variable and a latent factor (similar to a standardized regression coefficient)
- **Factor correlation（因子間相関）:** A “true” correlation between a pair of factors; adjusted for measurement error

Step 4: Interpreting key analysis results(2)

Communality （共通性）

Communality should be less than 1.0 for every single variable.

If communality > 1.0 for any, the factor solution cannot be interpreted in a meaningful way.

共通性

	初期	因子抽出後
Speak1	.394	.427
Speak2	.510	.547
Speak3	.528	.571
Speak4	.558	.622
Speak5	.535	.628
Speak6	.505	.553
Listen1	.415	.451
Listen2	.483	.545
Listen3	.264	.271
Listen4	.456	.517
Listen5	.421	.493
Listen6	.446	.510

因子抽出法: 主因子法

Step 4: Interpreting key analysis results(3)

Calculating communality

Example: When the inter-factor correlation (ϕ_{21}) = .66, the communality of Speaking 3 can be calculated as:

$$\text{Communality} = \lambda_{31}^2 + \lambda_{32}^2 + 2\lambda_{31}\,\phi_{21}\lambda_{32}$$
$$= (.70)^2 + (.08)^2 + 2(.70)(.66)(.08)$$
$$= .490 + .0064 + .074 = .570$$

(For details about calculating communality, see Brown, 2006, pp. 90-91)

Step 4: Interpreting key analysis results(4)

パターン行列[a]

	因子	
	1	2
Speak1	.558	.133
Speak2	.700	.057
Speak3	.700	.080
Speak4	.813	-.037
Speak5	.872	-.130
Speak6	.680	.092
Listen1	.074	.620
Listen2	.116	.656
Listen3	-.002	.522
Listen4	.060	.678
Listen5	-.031	.723
Listen6	-.066	.756

因子抽出法: 主因子法
回転法: Kaiserの正規化を伴うﾌﾟﾛﾏｯｸｽ法

a. 3 回の反復で回転が収束しました。

Factor loadings and inter-factor correlation after Promax rotation (based on PFA)

因子相関行列

因子	1	2
1	1.000	.663
2	.663	1.000

因子抽出法: 主因子法
回転法: Kaiserの正規化を伴うﾌﾟﾛﾏｯｸｽ法

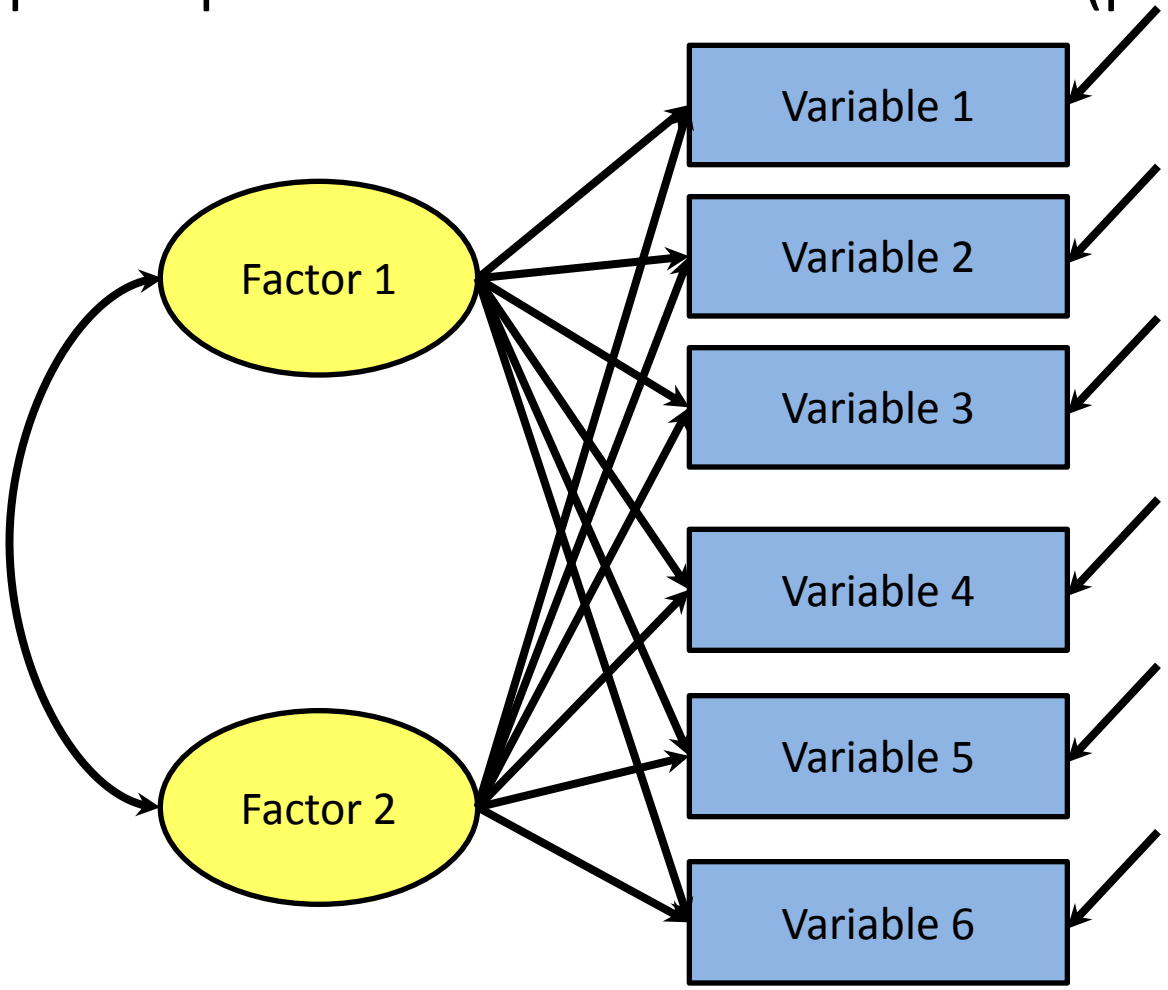
EFA vs. CFA (1)

Differences in approaches

- Exploratory factor analysis (EFA; 探索的因子分析)
  - Data-driven approach
  - Often used in early stages of an investigation
- Confirmatory factor analysis (CFA; 検証的/確認的因子分析)
  - Theory-driven approach
  - Often used in later stages of an investigation to confirm specific hypotheses

EFA vs. CFA (2)

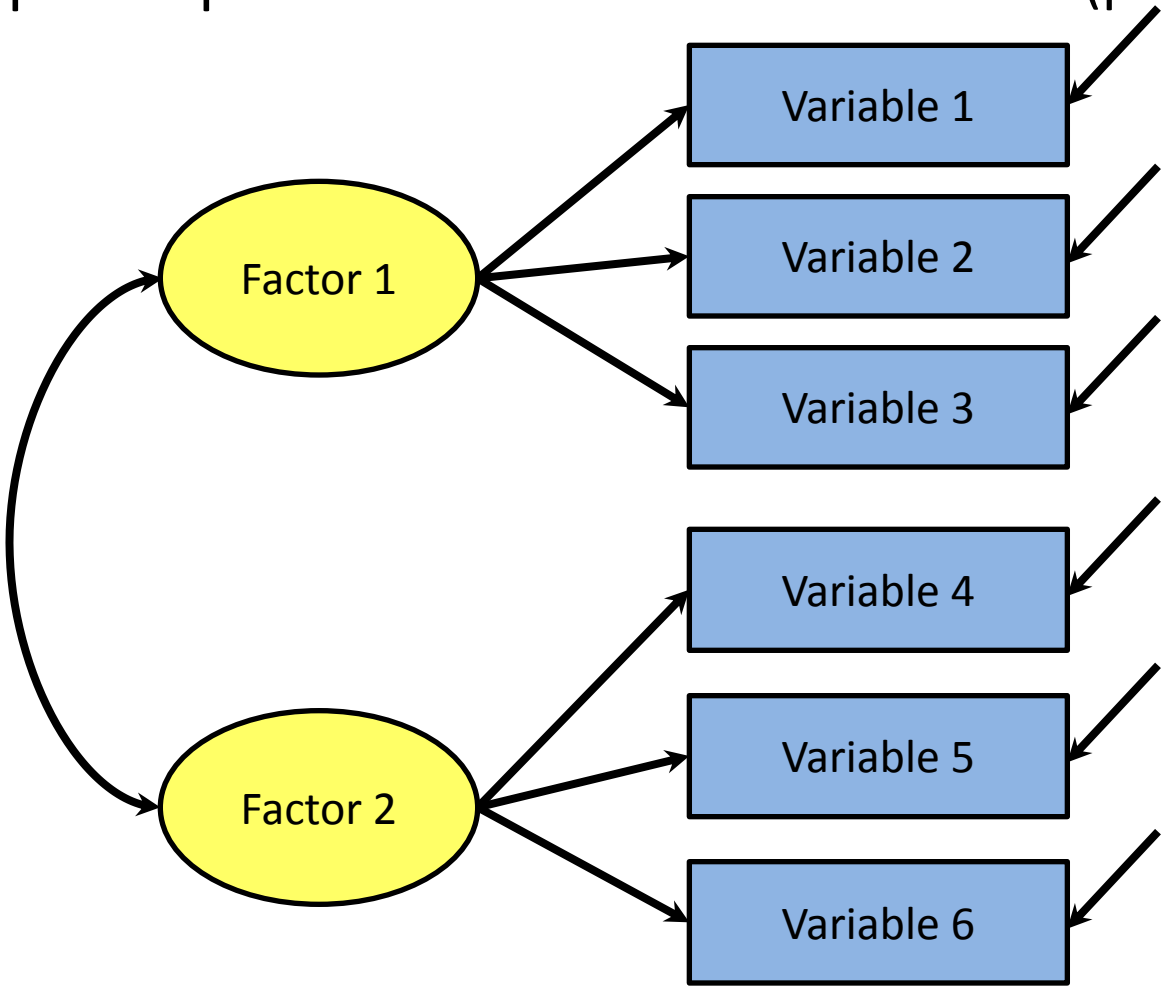
Graphic representation of the differences (per Brown, 2006)



EFA
(oblique rotation)

EFA vs. CFA (3)

Graphic representation of the differences (per Brown, 2006)



CFA

EFA vs. CFA (4)

Issues of consideration about using EFA and CFA (Jöreskog, 2007)

- Stages of research: exploratory vs. confirmatory phases
- Nature of investigation: "Factor analysis need not be strictly exploratory or strictly confirmatory. Most studies are to some extent both exploratory and confirmatory because they involve some variables of known and other variables of unknown composition." (Jöreskog, 2007, p. 58)
- Cross-validation of results from exploratory studies by conducting confirmatory analyses on different data sets

  (e.g., using randomly-split samples)

Useful guidelines for conducting EFA

- Factor analysis is often criticized not because of the fundamental weakness of the methodology but because of its misuses in previous factor analysis applications

- Further reading:
  - Brown (2006)
  - Fabriger et al. (1999)
  - Floyd & Widaman (1995)
  - Preacher & MacCallum (2003)
  - Tabachnick & Fidell (2007)

References

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language.* Cambridge, U.K.: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 26,* 9-48.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford.

Cattell, R. B. (1966). The scree test for the number of common factors. *Multivariate Behavioral Research, 1,* 245-276.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fabriger, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

References (continued)

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Jöreskog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions.* Mahwah, NJ: Erlbaum.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151.

Liu, O. L., & Rijmen, F. (2008). A modified procedure for parallel analysis of ordered categorical data. *Behavioral Research Methods, 40,* 556-562.

References (continued)

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*(1), 13-43.

Sawaki, Y. (in press). Factor analysis. In Carol A. Chapelle (Ed.), *Encyclopedia of applied linguistics.* New York: Wiley.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Vandergrift, L., Goh, C. C. M., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning, 56*(3), 431-462.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432-442.

Questions?

E-mail me at: ysawaki@waseda.jp