STATISTICAL HORIZONS

Scale Construction and Development

Tenko Raykov, Ph.D.

Upcoming Seminar:
April 26-27, 2019, Philadelphia, Pennsylvania

Plan of Course
(main topics covered):

1. **Factor analysis: A modeling basis of instrument construction and development.**

2. **Scale development with categorical items: Latent structure examination.**

3. **Construction of initial measuring instrument version.**

4. **Scale revision for enhancing psychometric quality.**

5. **Essential unidimensionality of multiple component measuring instruments.**

6. **Some practical issues in scale construction and development.**

7. **Instrument development with data from nationally representative studies.**

8. **Optimal shortening of psychometric scales.**

9. **Conclusion.**

1. Factor Analysis: A Modeling Basis of Instrument Construction and Development in the Social and Behavioral Sciences

1.1. Introduction

Scale construction and development (SCD) is a *multi-phase process* of major relevance in the social and behavioral sciences, and especially in empirical research.

The latter depends critically on *high quality data*, in particular those stemming from highly reliable and highly valid measuring instruments, such as scales, self-reports, tests, batteries, inventories, etc. (generally referred to as 'scales' or 'measuring instruments'/ 'instruments' in this course, as mentioned earlier).

SCD can be a rather *complicated process* in empirical work.

SCD depends crucially on available *substantive knowledge* in a subject-matter area of concern, and need not always proceed following exactly the same sequence of steps (or the same steps).

In this Section 1 of the course, we cover a main modeling framework that is very often used in SCD and particularly in its early phases. This is the framework of *factor analysis* (FA) that is based on the concept of *latent variable* (LV) defined shortly.

As in many empirical settings across the social and behavioral sciences working with already available instruments, *we assume throughout this Section* 1 *that a set of components is given comprising an initial (tentative) scale version*. Its features are of relevance to be examined, specifically its *underlying structure*, and if need be this scale is to be *improved* (see Section 3 on how to get to that scale).

Definition of latent variable

In SCD, a question of major relevance – especially in its initial stages (but not only) – is the following one:

> **If several items (measures, components, tests/subtests in a battery) are administered to a sample of subjects from a studied population, do these *items assess the same underlying construct, or perhaps more than one latent dimension (variable)*?**

In the latter case, one is also interested in knowing which of the measures are indicative of which latent variable.

Definition: A *latent variable* (LV, also referred to as 'construct', 'latent dimension', 'factor', 'trait', 'underlying variable', or 'hidden variable') is a random variable with the following properties:

- **it has *individual realizations* in a sample/population of interest,**
- **these individual realizations are however *not observed*.**

Examples: (particular) ability, competence, depression, intelligence, motivation, anxiety, cognitive functioning; attribute, aptitude, attitude, social cohesion, alienation; neuroticism, extraversion, acquiescence, openness to new experience, conscientiousness (The Big Five), etc.

With this definition in mind, we note that statistical modeling using latent variables is typically referred to as *latent variable modeling* (LVM). This course is developed within the LVM framework.

Counterparts of LVs are also found in the *hard sciences* (e.g., physics).

Back to the LVs: The LVs are only indirectly observable (measureable) random variables, which presumably have manifestations in observed behavior – e.g., answers to items or questions, responses to test components, etc.

Some LVs may *initially* be of *hypothetical* nature only and not very well defined. Empirical and theoretical research aims at improving their conceptualizations and more precise definitions.

The above major query (in red, at top of p. 9), is essentially the same as the one asking *whether there are one or more clusters of observed variables on which subjects display similar (related) performance*.

Thereby, within each cluster of measures, subjects give rise to markedly correlated scores, whereas across clusters (if the latter are more than one) their relationships are notably weaker.

For example, if a battery of intelligence tests is given to a group of subjects, it would oftentimes be of concern to know whether the *pattern of observed correlations* among these measures supports the hypothesis that there are say two different clusters of measures.

One of these clusters may pertain to the so-called fluid intelligence, the other to crystallized intelligence (e.g., Horn, 1982, *Handbook of Developmental Psychology*; they have been at times also alternatively referred to as performance and verbal intelligence, respectively).

This type of queries is appropriately addressed via use of *factor analysis* (FA). FA is in many aspects *the fundament of this course* and the vast majority of scale development in empirical research.

FA has its beginnings around the turn of the 20th century when the English psychologist Charles Spearman proposed it as a method of investigating his *bi-factor theory* (Spearman, 1904, *Amer. Psychol.*).

He maintained that there was a general intelligence factor, called *g*, which was involved in human performance on any intelligence test. In addition, each intelligence measure was characterized by a 'unique factor' that was specific only to that measure. Hence the name *bi-factor theory of intelligence* (also known as '*g-theory*').

1.2. The (classical linear) factor analysis model

A *statistical model* is typically defined as (i) a set of equations, *and* (ii) associated distributional assumptions about stochastic elements that are essential parts of the model.

Unlike the widely used statistical technique of *principal component analysis* (PCA, especially in the 'hard' sciences; Raykov & Marcoulides, 2008, fully cited earlier), *FA is based on a statistical model and has inherent residual terms* (see for instance chapters 8-11 of the last cited source, for similarities and especially distinctions between PCA and FA, to avoid confusion in empirical research).

The *model of* (classical) FA relates each observed, or manifest, measure to each of the assumed latent variables. Specifically, this model consists of as many equations as there are manifest measures.

With p observed variables, $X_1, X_2, \ldots, X_p$ ($p > 1$) and m factors, $f_1, f_2, \ldots, f_m$, *the* (*classical*) *FA model is as follows* ($1 \leq m \leq p$):

(1.1)

$$X_1 = \mu_1 + a_{11} f_1 + a_{12} f_2 + \ldots + a_{1m} f_m + u_1$$

$$X_2 = \mu_2 + a_{21} f_1 + a_{22} f_2 + \ldots + a_{2m} f_m + u_2$$

$$\ldots$$

$$X_p = \mu_p + a_{p1} f_1 + a_{p2} f_2 + \ldots + a_{pm} f_m + u_p ,$$

where $u_1, u_2, \ldots, u_p$ are so-called *unique factors* with zero mean, which are assumed uncorrelated among themselves and with the factors, the a's are referred to as *factor loadings*, and the μ's are observed measure intercepts (not of relevance in the remainder).

We also assume in this course that the factors have zero means, and that the *X*s are 'fixed' rather than randomly sampled from a pool of measures under consideration.

Further, we typically want that *m* be much smaller than *p*, i.e., the number of factors is at least considerably less than that of observed variables. This is of special relevance when using FA as a means of data reduction (recall this particular aim also of PCA).

The *unique factors* in Equation (1.1), $u_1, u_2, \ldots, u_p$, comprise (a) error of measurement as well as (b) all sources of variance in the pertinent observed variables ($X_1, X_2, \ldots, X_p$), which are not captured by the *m* factors, i.e., are <u>not</u> shared with the other observed measures. For this reason, the unique factors are also often referred to as *residual (error) terms*. That is, beyond (a) above, any unique factor represents the sources of all variance in the observed variable that it pertains to, which does not stem from the latent factors.

This is the reason why the latent factors, $f_1, f_2, \ldots, f_m$, are also referred to as *common factors*, and Equations (1.1) as the (classical) *common factor analysis model*. (Note that also the distributional assumptions mentioned after Equations (1.1) belong to the model.)

The *common factors*, $f_1, f_2, \ldots, f_m$, can be viewed as the sources of all *shared (common) variability* among the *p* observed variables. We will continue to use however the popular reference 'latent factors' ('factors' or 'traits') for the *f*'s in the rest of this course.

Note that strictly speaking the unique factors are also latent variables, since they are <u>un</u>observed. However, we will reserve the name 'latent factors' or 'factors' only for the common factors in this course (i.e., for the 'proper' factors, $f_1, f_2, \ldots, f_m$, $m > 0$).

Throughout the course, we will also assume that each considered latent variable (factor, dimension, trait, ability, construct) is *continuous*.

Alternatively, I note in passing that *latent class analysis* – which is not the subject of this course (but alternative ones) – considers the latent variables involved as categorical (or mixtures, generally).

I stress that the FA model in Eq. (1.1) has the assumption that *none of the unique factors is related to any other unique factor or to a common factor* – the former part of this assumption can be relaxed in a different mode of FA (as we will do later in this Section 1).

Sometimes, in addition to all above assumptions, one also makes that of *normality* of the observed variables. This assumption will be relevant when interested in hypothesis testing, or applying a particular method of FA, specifically using the popular maximum likelihood (ML) method for the purpose of 'factor extraction' (more on this matter comes later in the course).

There is a clear (conceptual) analogy between the FA model in Equations (1.1), on the one hand, and that of regression analysis.

We will keep in mind this analogy throughout the course, but stress that while Equations (1.1) resemble strongly a multivariate multiple 'regression analysis' model (general linear model, abbr. GLM), in empirical research we cannot really carry out that regression since the factors are not observed and hence there are no data on them.

In this analogy context, one sees that parameters of the discussed FA model are the following:

(1) the *factor loadings*, as counterparts of the partial regression coefficients whose role they play in Equations (1.1);
(2) the *variances of the unique factors*, as counterparts of the model error variance (squared standard error of estimate, in case of univariate multiple regression); and
(3) *factor correlations* (covariances).

It will be quite beneficial to be aware of these parameters when conducting FA for the purposes of SCD (as we will do in this course) or for other aims. This is because one knows well a given model only when he/she knows well (a) which its *parameters* are, and (b) the *assumptions* underlying it.

A succinct representation of the FA model in Equations (1.1) is as follows:

(1.2) $$\underline{X} = \underline{\mu} + A\,\underline{f} + \underline{u}\ .$$

In Equation (1.2), $A = [a_{jk}]$ is the matrix of factor loadings while $\underline{X} = (X_1, X_2, \ldots, X_p)'$, $\underline{f} = (f_1, f_2, \ldots, f_m)'$, $\underline{u} = (u_1, u_2, \ldots, u_p)'$ and $\underline{\mu}$ are the vectors of observed variables, common factors, unique factors, and intercepts, respectively. (In this course/booklet, brackets are used to enclose matrix elements and parentheses for vector elements, with priming denoting transposition and underlining vector).

From Equation (1.2), using the well-known rules for working out variances and covariances of linear combinations of random variables (e.g., Raykov & Marcoulides, 2006, fully cited earlier), one obtains in *compact matrix terms* the following expression:

(1.3) $$\Sigma_{xx} = A\,\Phi\,A' + \Theta\ .$$

In Equation (1.3), which is a very popular representation of the *implications* of the FA model in Equations (1.1) (and the assumptions in the discussion immediately following it above, pp. 11-12), Σ_{xx} is the covariance matrix of the observed variables, Φ is that of the latent factors (which need not be diagonal), and Θ is the covariance matrix of the unique factors.

The error covariance matrix, Θ, is *assumed diagonal* - in this exploratory treatment of FA (see next subsection for extension) - as indicated earlier. Also, a similar decomposition to Equation (1.3) holds for the correlation matrix of the observed variables.

With the above in mind, the *critical question of FA* as a statistical modeling and analysis method, is the following:

> **What is the minimal number *m* of latent factors, such that given (i.e., fixing/conditioning on) the *m* factors, the observed *p* variables are uncorrelated (independent)?**

Note that this statement does not depend on the scale of the observed variables. It implies that it is the *common factors* that comprise then all *sources of relationships* among the observed measures, and hence in this sense 'explain' their relationships.

Further statistical details of this fundamental, critical question and issue for many applied statistical methods – such as IRT, latent class analysis, latent profile analysis, and factor analysis – can be found in Bartholomew, Knott, & Moustaki, 2011, *Latent variable models and factor analysis: A unified approach*, London, UK: Arnold.

In fact, it is this major question (in red above on this page), which *unifies* and ties all these methods.

A useful diagram notation (pictorial representation of LV models)

In SCD, like in many other applications of FA, it will be very helpful to be in a position to graphically represent models of interest.

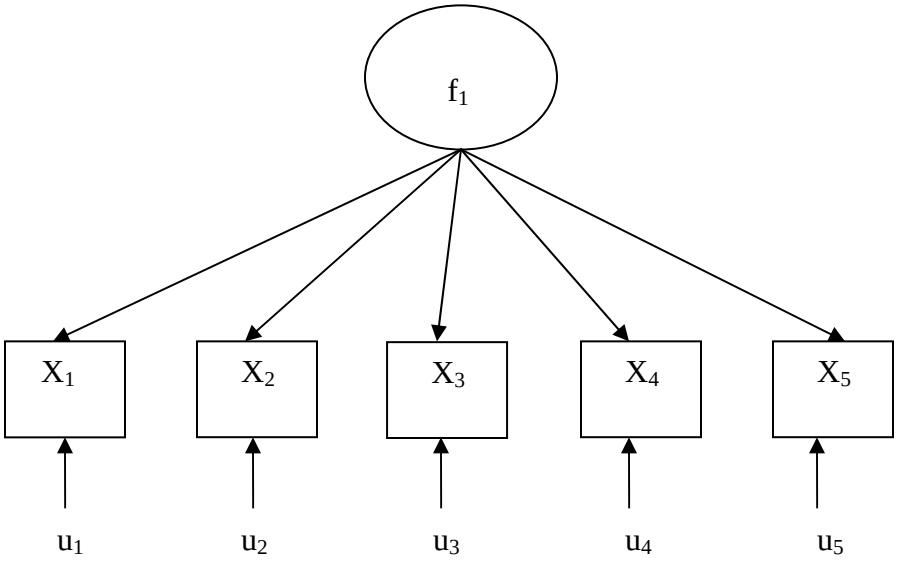
We can use for this purpose the popular *path-diagram notation*.

In particular, suppose we place the factor(s) in *circles* (or ellipses), and the observed measures in *squares* (rectangles).

Further, let's use a *1-way arrow* to represent the assumption that the factor plays explanatory role for a given observed measure, and a *two-way arrow* to represent a covariance (correlation), without any specific assumption of 'causal' relationship between the variables involved.

Then a single-factor model (i.e., a unidimensional model) with five observed measures looks pictorially as follows.

***Figure* 1.1**



Observe that this is a graphical representation of a homogeneous/ *unidimensional scale* with 5 components.

Such scales, i.e., unidimensional/homogeneous measuring instruments are widely sought in empirical social and behavioral research. With them, the observed measures evaluate (presumed) underlying single latent variables that can be, and typically will be, of main research interest.

I should like to point out here that in principle a multiple-construct scale *can* have potential issues when being used (as is) in empirical work, as we discuss next. (More on this matter comes also later in this Section 1 and in the course – e.g., Sections 5 and 6.)

Why are unidimensional scales (subscales) of such interest in theoretical and empirical research?

The reason is that a high (low) score on the observed variables (or their sum, weighted or not) cannot be <u>un</u>ambiguously attributed to high/low score on any of the underlying latent variables.

The situation is even more problematic when trying to interpret substantively in a similar manner medium large values on the overall scale.

This possibly serious *interpretational difficulty* does <u>not</u> arise with homogeneous (unidimensional) scales. That fact is a major reason why the latter are so widely sought in empirical research.

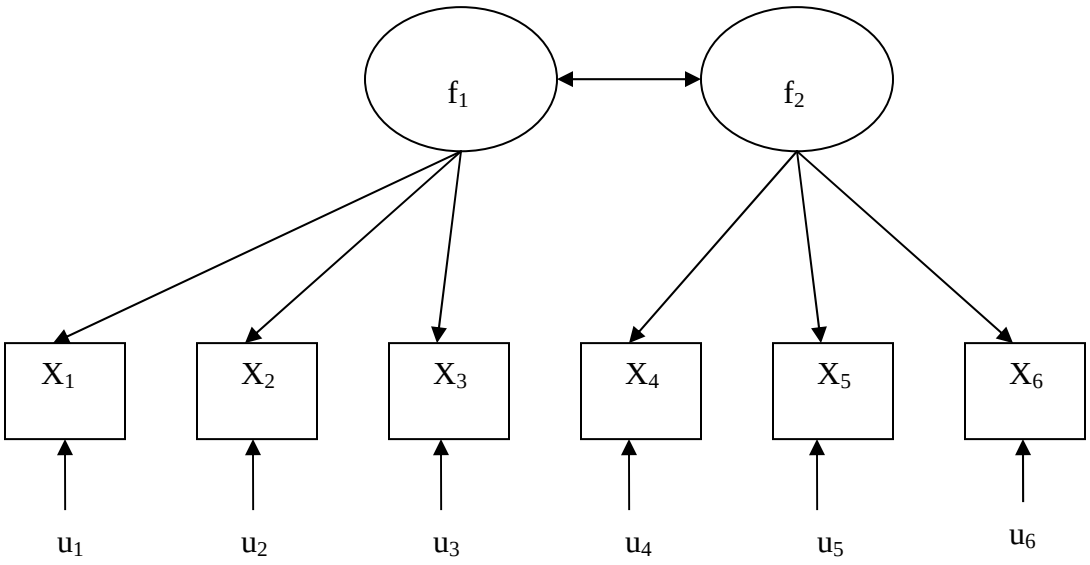
I would like to stress, however, that it is (expert) *substantive considerations* and judgment, which should govern a decision whether to use a multi-construct (multidimensional) scale or not, since *the bottom line of behavioral and social measurement is <u>validity</u>*.

In particular, it is known that some scales inherently measure more than 1 dimension and still can be rather useful and considered valid measuring instruments in research (e.g., of depression, self-esteem).

We address further aspects of this important issue later in the course, e.g., when considering statistical methods that can help one decide if a scale is essentially unidimensional.

Similarly, *if an initially considered scale is multidimensional, it may be possible to build ('break it down' to) subscales from it that are each unidimensional*. This is also a main point of the present section that is worth keeping in mind.

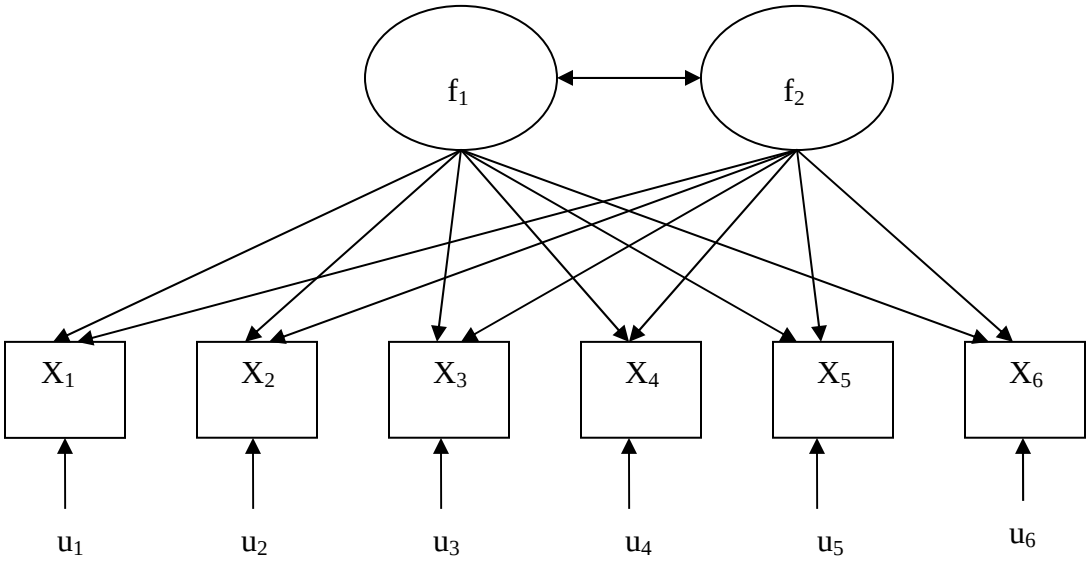
An example is given with the model depicted in the figure displayed next.



Back to our earlier discussion of utility of the widely followed path-diagram notation, I should like to point out that a two-factor model with six observed measures loading on both, is presented pictorially as follows in Figure 1.2, whereby a two-way arrow denotes correlation.

Observe that this is the diagram of a two-factor scale that is *not* homogeneous/unidimensional. Rather, the scale next evaluates or assesses a pair of latent variables, denoted f_1 and f_2. (Here, we are not concerned with parameter estimation/model identification – as we will be later.)

***Figure* 1.2**



It is highly recommendable that the path-diagram notation be used as often as possible, whenever one has an idea of (a) how many factors there could be behind a given set of observed measures, e.g., a scale under consideration, and (b) which of these manifest (observed) variables load on which factor.