STATISTICAL HORIZONS

Item Response Theory

Tenko Raykov, Ph.D.

Upcoming Seminar:
October 26-27, 2018, Philadelphia, Pennsylvania

Item Response Theory

Tenko Raykov
Michigan State University
(www.msu.edu/~raykov)

Citation of this booklet:

Raykov, T. (2017). Item Response Theory. Course Booklet. Statistical Horizons, Philadelphia, PA.

Plan:

0. **Resources for course.**
1. **Introduction and overview of Item Response Theory (IRT)/Item Response Modeling (IRM). The logistic function and the normal ogive, and getting to know Stata.**
2. **A start-up example of IRT/IRM.**
3. **Popular unidimensional IRT models.**
4. **Parameter estimation in item response models.**
5. **Item information and test information functions. Test characteristic curves. Scoring of studied subjects.**
6. **Measuring instrument construction, development, and revision using IRT/IRM.**
7. **Differential item functioning and methods for its examination.**
8. **Polytomous IRT models.**
9. **Multidimensional IRT/IRM.**
10. **Extensions and limitations of some current IRT/IRM applications.**
11. **Conclusion and outlook.**

0. Resources for Course

- ***Literature*:**

Raykov, T., & Marcoulides, G. A. (2017). *Item response theory and modeling using Stata.* College Station, TX: StataPress.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Factor analysis and latent variable models*. New York: Wiley.

Lord, F. M. (1980). *Item response theory and its applications.* Hillsdale, NJ: Erlbaum.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York, NY: Taylor & Francis.

Raykov, T., & Calantone, R. J. (2014). The utility of item response modeling in marketing research. *Journal of the Academy of Marketing Science, 42,* 337-360.

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement, 76,* 325-338.

Van der Linden, W. J. (2016) (Ed.) *Handbook of Item Response Theory.* Boca Raton, FL: CRC Press (Chapman & Hall/Taylor & Francis).

- ***Software:***

Stata – module 'irt' (v. 14; www.stata.com).

Mplus – Muthén & Muthén (2017). *Mplus user's manual*. Los Angeles, CA: Muthén & Muthén (www.statmodel.com).

I will occasionally also refer to the packages flexMIRT, IRTPRO (www.vpgcentral.com; www.ssicentral.com) and R's package 'ltm'.

***Notes and Disclaimers*:**

0. This document provides *more* material and information on its topics and about IRT/IRM generally, than I will be able to cover (verbally) during the workshop. Therefore, I recommend you taking a closer look at it at a later time, to maximize the benefits from this short course.

1. The data sets used are only employed for the purpose of method *illustration*, rather than to reach any deep (or new) substantive conclusions in their subject-matter domain.

2. Due to the time constraints of the course, we will *not* be able to use *all* packages on all examples (see also Appendices to some of the following sections).

3. Please treat this booklet as containing *copyrighted* material only, which is provided just for your use, and do not make it available to anyone else in any form or on any occasion. Thank you.

4. On some of the pages that follow, there may be some blank lines around the middle – they're used solely for *pagination*/slide presentation purposes (i.e., nothing has been deleted there).

5. In the remainder, in the context of *binary* or binary scored items (as with Likert-type items that are binary scored in the end), I will generically use the term '*correct*' response for an answer by a studied subject that is 'true', 'endorsing', 'present', 'agree', 'positive' 'applies to me', 'successful

outcome' ('success') etc.; and to the other possible answer(s) as 'incorrect' (which is for 'fail', 'negative', 'false', 'not present', 'not endorsing', 'does not apply to me', etc.); these answers are denoted '1' and '0', resp.

Similarly, we'll make the standard statistics' *independence* assump-tion that observations are unrelated to each other. (Extensions later.)

6. **Although we employ throughout Stata and/or Mplus, as well as make references to 3 other software, this short course is *not* about IRT software. Rather, this w/shop is about IRT/IRM itself, and illustrates it with particular software use.**

7. **The workshop can be seen as being characterized by its *features* of (i) being free of (a) incorrect statements about, or treatment of, classical test theory (CTT) and (b) the CTT juxtaposition to IRT/IRM; as well as (ii) stimulating latent variable modeling (LVM) based 'thinking' and in particular LVM-based understanding of IRT/IRM. The workshop therefore empowers additionally the empirical scientist using IRT/IRM to capitalize on this *special connection*:**

1. Introduction and Overview of Item Response Theory

Item response theory **(IRT; item response modeling, IRM) is an applied statistics and measurement discipline concerned with mathematical and stochastic functions describing**

(i) the *interaction of examined persons and a measuring instrument and its components*, such as items, testlets, subtests, subscales, questions, etc.; as well as

(ii) the information contained in the data from the instrument with respect to its components (referred to as 'items' below), their functioning, overall instrument functioning, and studied persons.

Fundamental for IRM/IRT is the relationship, i.e., *function*, between these *two key concepts*:

(a) a *latent dimension* (trait, construct, ability, aptitude) being evaluated,
usually denoted θ ('theta'; may be uni- or multi-dimensional); and
(b) the *probability of 'correct'* (*particular*) *response* on a given item (presumably) assessing that trait/ability, as a *function* of θ.

Each such function, denoted $P(\theta)$, which describes this relationship is called *item response curve* (ICC), and is typically a monotone increasing function of θ.

After its class (or classes) is chosen by the researcher, this function is postulated usually for *each* item, with *different* parameters in general across items.

Other frequent references to the ICC that can be found in the literature are *item characteristic function*, *item response function, item trace curve*, *or item trace line* (with ICC typically used, and in this course).

The concept of ICC is so *fundamental* for IRT that actually as a *single sentence definition* of IRT we can give this one:

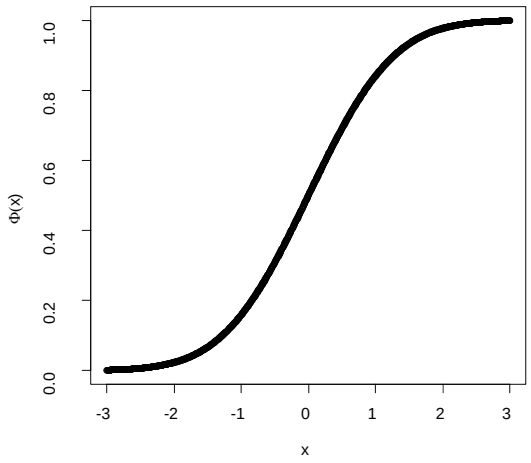
> **IRT = applied statistical framework for modeling the ICC, or $P(\theta)$, for each item in a measuring instrument or item set under consideration.**

There are *infinitely* many possible choices for ICC as a function of θ, i.e., for $P(\theta)$, and *two* of them – mentioned next – have obtained special and nearly exclusive prominence in IRT/IRM.

One of these functions, the *normal ogive*, is displayed below and is the cumulative distribution function (CDF) of the standard normal distribution, often denoted $\Phi(x)$ (see, e.g., Figure 1.1).

The other is the CDF of the *standard logistic* distribution, denoted $\Lambda(x)$, which (a) has the same shape, and (b) under an appropriate scaling (unit change) differs from $\Phi(x)$ by less than .01 in terms of response probability across the entire real line (i.e., for any x; see below for their difference). We discuss it in more detail in a later subsection of this section.

***Figure* 1.1**



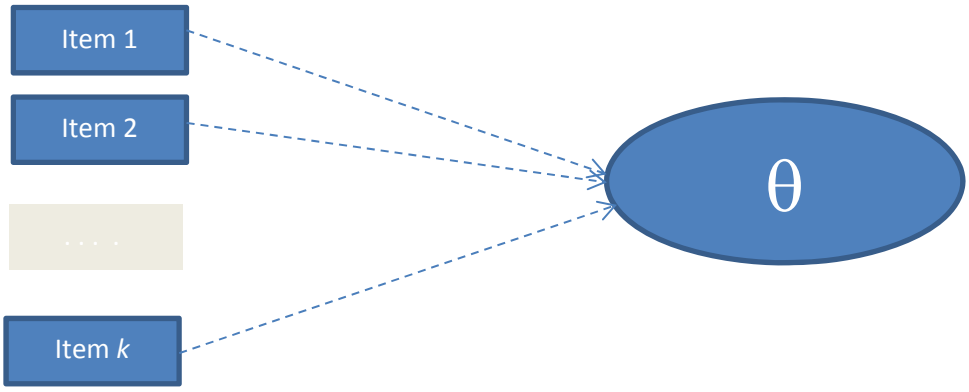
As indicated above, IRT/IRM is concerned with *measuring instruments* consisting of multiple components (referred to as 'items' in this course).

They provide *multiple converging pieces of information about underlying unobserved traits, constructs, abilities, attitudes*, in general presumed *latent dimensions* (often referred to as 'traits', 'constructs', 'abilities', 'continua', or just 'latent dimensions' below), which are denoted θ and as mentioned may be uni- or multi-dimensional, i.e., scalar or vector variables.

While θ is/are <u>un</u>observed, (presumed) manifestations of it/them are typically the *responses* obtained from the studied/examined subjects or respondents on the components of the measuring instruments, which are items, questions, tasks, problems to solve, or in general the instrument elements (generically referred to as 'items' in the remainder, as indicated above).

A simple schematic *representation of the aims* of IRT/IRM is given in Figure 1.2 next (this representation is *not* meant to be any kind of model and is used here *only* for conceptual purposes, to emphasize we wish to make inferences about θ using the *k* items; *k* > 1; for further details, see Section 2).

***Figure* 1.2**



The *connection* between the individual items and θ is facilitated as alluded to above by the pertinent assumed ICC, i.e., *P*(θ). This function *P*(θ) usually belongs to the same functional class for all items, or to 2 or more classes as in hybrid IRT models (see, e.g., Raykov & Marcoulides, 2017, and later in w/shop).

With the above in mind, it's worth pointing out that in this course *we will not be concerned* with timed or speeded tests or time-limited behavior measuring procedures, or IRM with continuous indicators – see, e.g., van der Linden, 2016, for details on this kind of tests that are currently infrequently used in the behavioral, social, educational, biomedical, clinical, nursing, communication, or marketing sciences.

So, what is the IRT/IRM process in a nutshell?

IRT/IRM essentially deals with

(a) postulating models about that relationship, which instrumentally involve unknown parameters, for a set of *k* items or given measuring instrument (*k* > 1, as assumed throughout this workshop);

(b) estimating these models using an available data set and carrying out model choice; and

(c) based on (a) and (b), estimating (predicting) individual subject values for θ.

A particularly important feature of IRT/IRM is that at the end of its application, using a plausible model for an available data set from a given measuring instrument, the researcher obtains the following two *sets of quantities that are commensurate* (i.e., are associated with/are on the same continuum):

(i) a set of quantities/parameters (estimates) characterizing the items,

and

(ii) a set of quantities/values (estimates – one per person, in unidimensional IRT) characterizing the extent to which each person possesses the trait(s) evaluated.

How to accomplish the IRT/IRM process optimally, is the concern of the remainder of this workshop.

To get us on the way to specific discussions of IRT/IRM, we will discuss next in more detail with the normal ogive and especially the logistic function that's fundamental for the remainder of the course, and will get thereby acquainted with Stata as well.

The logistic function and the normal ogive

As we indicated earlier, in IRT two main functions are used for modeling each item's ICC, $P(\theta)$ - these are the normal ogive (which is actually rarely utilized these days), and the logistic function.

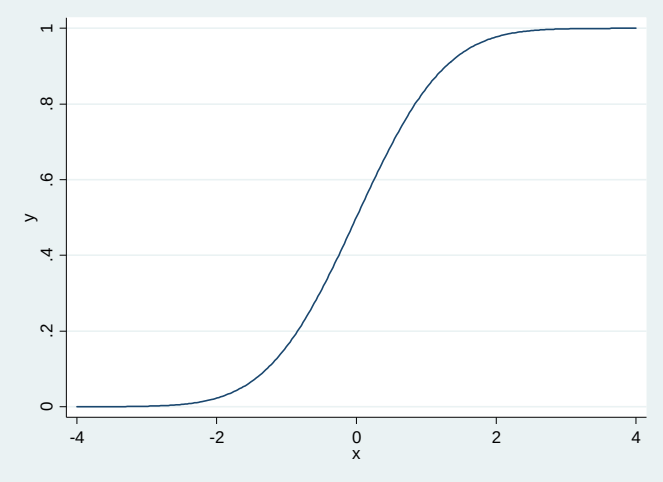
The *normal ogive* is as mentioned the cumulative distribution function (CDF) of the standard normal distribution, denoted $\Phi(x)$ (see Figure 1.1 above, or Figure 1.3 given next). The normal ogive represents the standard normal probabilities, i.e., the probabilities of picking at random a score that is less than *x* (for any *x*), from a standard normal population distribution or random variable following the standard normal distribution.

We can readily get a plot of the normal ogive with Stata this way (command preceded by the Stata prompt, which is a dot and you need not type in Stata's Command window; see, e.g., Raykov & Marcoulides, 2017, for a more detailed discussion):

```
. twoway function normal(x), range(-4 4)
```

The result of this command is presented in Figure 1.3, which is in fact displaying a prototypical ICC.

***Figure* 1.3**



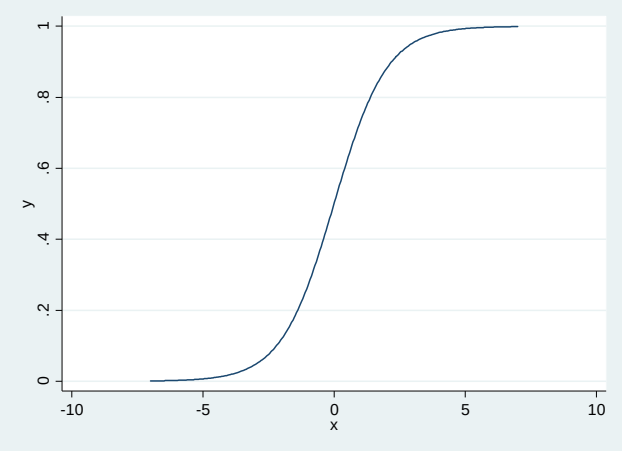
The other function with truly special relevance in contemporary IRT/IRM and also this workshop, is the CDF of the *standard logistic* distribution (referred to as 'logistic' or 'logistic function' for simplicity), which is denoted $\Lambda(x)$.

The logistic function, or logistic curve, (a) has the same shape, as well as (b) under an appropriate scaling (unit change) differs from $\Phi(x)$ by less than .01 in terms of response probability across the entire real line (i.e., for any *x*; see below for a more specific discussion of their difference).

We can readily obtain with Stata the logistic function, or logistic curve (see Raykov & Marcoulides, 2017, for a more detailed discussion, and Figure 1.4 below):

```
. twoway function logistic(x), range(-7 7)
```

***Figure* 1.4**



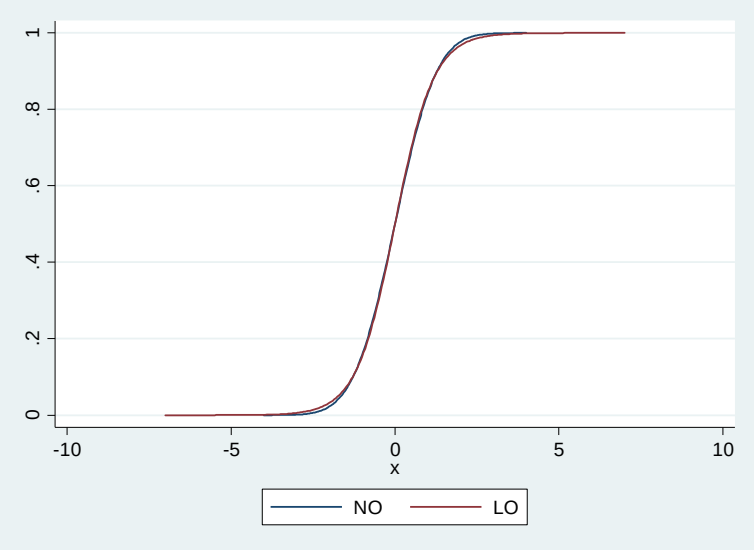
In fact, the relevance of the logistic function (curve) for IRT/IRM is so high, one could even say that *if there was a single most important function in present-day IRT/IRM applications, it is the logistic*.

Note the distinct elongated S-shape of the logistic in the above Figure 1.4, which is very similar to that on Figure 1.1, i.e., of the normal ogive.

In actual fact, the normal ogive and the logistic function are essentially identical for most practical purposes (if not nearly all), as we can see this way (see Figure 1.5 below, and Raykov & Marcoulides, 2017, for more detail):

```
. twoway function NO = normal(x), range(-4 4)||function
LO = logistic(1.701*x), range(-7 7)
```

***Figure* 1.5**



As indicated above, the discrepancy between the two curves is less than .01 anywhere on the horizontal axis after this re-scaling is done for the logistic, which has no practical implications or substantive meaning (Healy, 1952).

For this reason, in addition to the mathematical and numerical tractability in particular of the logistic function, contemporary IRT applications use routinely - as in this w/shop - IRT models based on this function, which are referred to as *logistic IRT models*.

Due to this special importance of the logisitic function, denoted as mentioned throughout this booklet as $\Lambda(x)$, here's its formal definition:

(1.1) $$\Lambda(x) = e^x/(1+ e^x) = exp(x)/[1+exp(x)],$$

where the frequently used notation $exp(x)$ stands for e^x.