POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS

# Reflections on Big Data influence on The Structural Equation Modeling

Piotr Tarka

# Some facts related with BD

➢ Today’s society is more than ever connected with the information via Internet

- In 2000, 25% of the world's stored information was digital. More than 98% of all stored information is digital
- The digital universe is projected to explode to around 44 trillion gigabytes in 2020 (Turner, Gantz, Reinsel & Minton, 2014).

**Consequences** (Kaisler et al. 2013; Elgendy and Elragal, 2014):

1. Appearance of ultra-fast global IT connections; electronic databases and information systems, leads to overwhelming amount of data.
2. It descends on many communities, from governments, education and e-commerce, business to even health organizations.
3. BD changes the way we look at science (e.g. Finucane, Martinez & Cody, 2018)

# BD and Science

## Any consequences for scientists?

1. Scientists caught in a data deluge with unprecedented **volumes of information**,
2. To measure the social, behavioral and psychological phenomena, sholars need to adopt new analytical strategies and obtain knowledge from other disciplines to control BD.

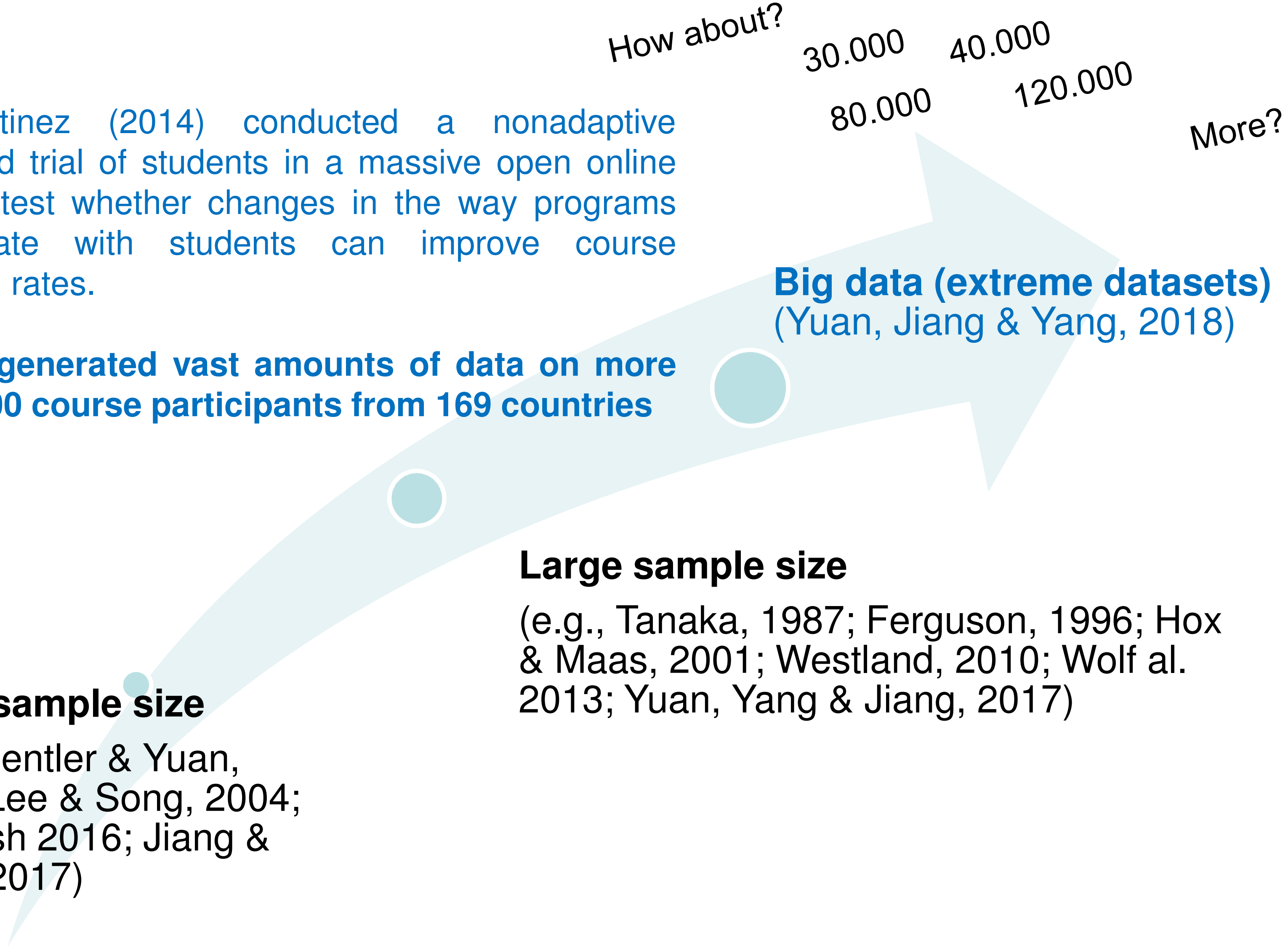
Note: relativeness of the „big data" as the term.

Sayeed Choudhury from John Hopkins University, Maryland, US, said that (…) „we should not just look at volume of data, we should also look at methods. Big data is when the method breaks down, when we need a completely new method to analyze the data that you have available."

## The question is what big data is (large datasets: observations, variables)?

e.g., Martinez (2014) conducted a nonadaptive randomized trial of students in a massive open online course to test whether changes in the way programs communicate with students can improve course completion rates.

**The RCT generated vast amounts of data on more than 23,000 course participants from 169 countries**



How about?

30.000 40.000

80.000 120.000

More?

**Big data (extreme datasets)**
(Yuan, Jiang & Yang, 2018)

**Large sample size**

(e.g., Tanaka, 1987; Ferguson, 1996; Hox & Maas, 2001; Westland, 2010; Wolf al. 2013; Yuan, Yang & Jiang, 2017)

**Small sample size**

(e.g., Bentler & Yuan, 1992; Lee & Song, 2004; McNeish 2016; Jiang & Yuan, 2017)

# Structural Equation Modeling – quick reminder

- A specific **theory-based causal connections** between latent variables and between those latents and relevant indicator variables

- Estimates of the model’s parameters represent values and imply the **variance/covariance matrix** that should be as similar as possible to the **data variance/covariance matrix**.

- **The model implied covariance matrix would be the population covariance matrix if the model was the proper model**.

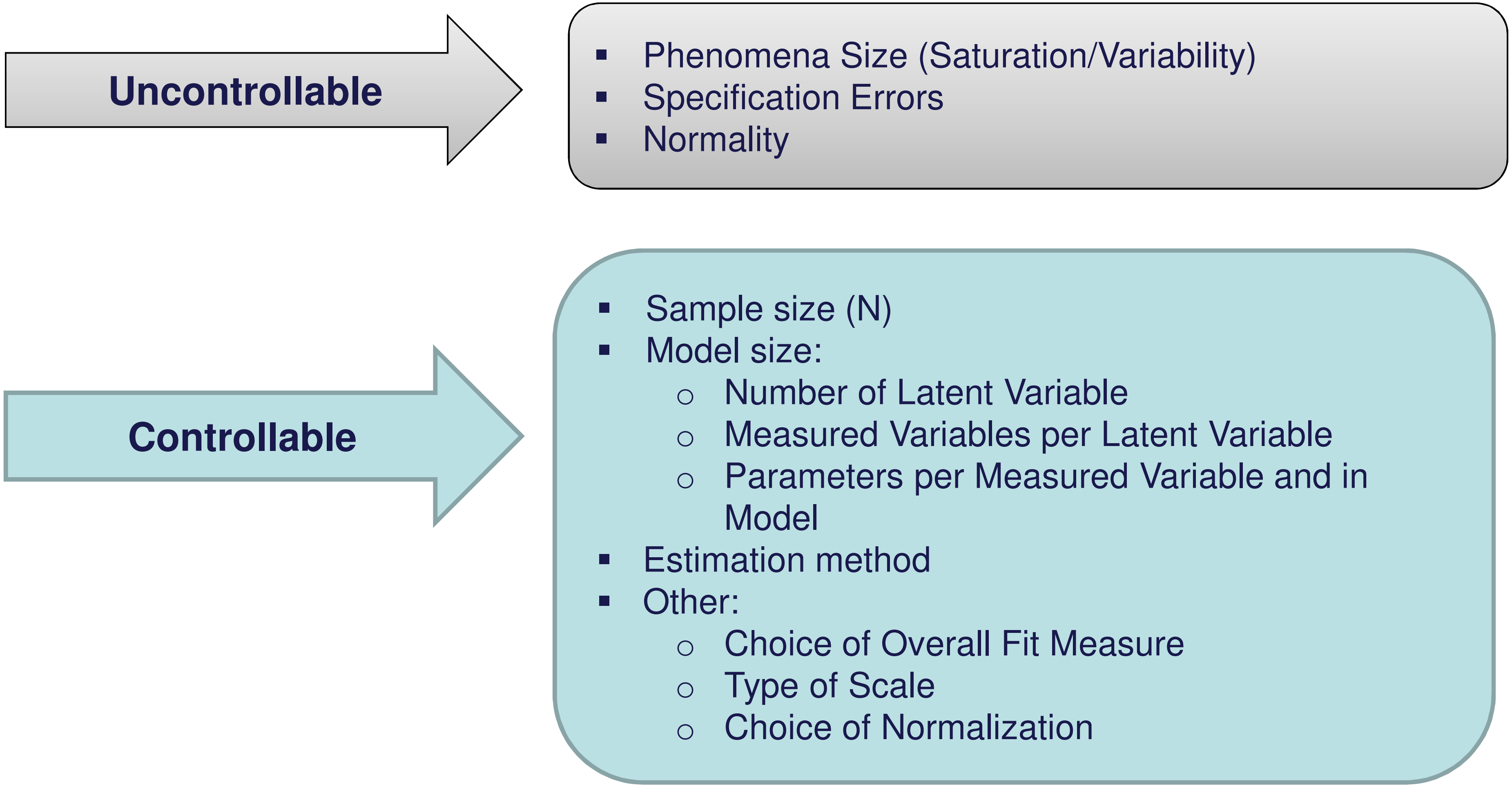
In other words, we base our current understanding of ‘‘how the world works’’ on SEM models, and use the diagnostic evidence accompanying to find out whether they fit or not (Hayduk, et al. 2007).

# Supporting rules for confirmation of the SEM model

- The measurement model is strong (e.g., 3 or 4 indicators per factor, and good reliabilities)
- Few and significant parameters in model
- **The structural path model is not overly complex or misspecified**
- **Variables reflect normality**

**The problem is that, in most of the social, psychological, behavioral or genetic scientific research projects, data often break the above rules** (West, Finch, & Curran, 1995; Anderson, 1996; Micceri, 1989; Blanca et al. 2013; Nicolaou & Masoner, 2013; Cain, Zhang, & Yuan, 2017).

# Controllable and uncontrollable SEM characteristics



**Uncontrollable**

- Phenomena Size (Saturation/Variability)
- Specification Errors
- Normality

**Controllable**

- Sample size (N)
- Model size:
  - Number of Latent Variable
  - Measured Variables per Latent Variable
  - Parameters per Measured Variable and in Model
- Estimation method
- Other:
  - Choice of Overall Fit Measure
  - Type of Scale
  - Choice of Normalization

Source: Nicolaou & Masoner (2013)

# Specification problems in SEM

Hayduk, et al. (2007) argued that:

- **Proper model/theoretical specifications** should imply covariance matrices that are within **sampling fluctuations of the data**,

- ....however, even if the model is **properly specified and the estimates provide proper parameter values, random sampling fluctuations can still keep the data matrix from corresponding exactly to the model-implied covariance matrix** (degree of ill fit).

- Differences between a model's implications and the data might not result from mere chance sampling fluctuations, but **misspecification that originates in real theory deficiencies.**

# The measurement model is strong (3 - 4 indicators per factor, good reliabilities and normality)

- **Two variables loading on a factor = bias in the parameter estimates**, but with three or more indicators, this bias nearly vanishes (Gerbing & Anderson 1985, p.268).
- **Strong, clean measures** are compensatory for sample size,...and the number of variables per factor may have an effect on improving fit statistics (Jackson, 2003).

Nonrobust parameters, weak variables lead to alternative approaches: **asymptotically distribution-free methods, bootstrapping**, or **nonparametric methods**, **but these usually need large sample sizes** (at least 3000-5000, Finch, West, MacKinnon, 1997).

# Few parameters in SEM

- Models with numerous parameters = **probability that any parameter will be significant by chance increases as a function of the number of parameters to be tested in the model** (Cribbie, 1999).

- **As number of parameters to be evaluated increases, there increases the probability of <u>falsely</u> declaring individual parameters significant, and <u>falsely</u> declaring relationships significant in the model.**

Finally, there is a high degree of interrelatedness between parameters in a model (Kaplan & Wenger, 1993).

# Sample Size and Model Size in SEM – threats or opportunities?

- **Number of indicators *(p)* per latent variable *(m)*** (e.g., Gerbing & Anderson, 1985; Marsh et al., 1998; Velicer & Fava, 1998).

- **Number of indicators *(p)* exceeding number of observations *(N)*** (e.g., Ferguson, 1996; Kenny & McCoach, 2003; Yuan, Yang & Jiang, 2017; Yuan, Jiang & Yang, 2018)

- **Excessive number of parameters *(q)*** (e.g., Bandalos, 1997; Boomsma, 1982; Gerbing & Anderson, 1985; MacCallum et al., 1999; Velicer & Fava, 1998; Cribbie, 2000).

# Positive effects in SEM

- When there are more latent variables in a model,
- When there are more measured variables per latent variable,
- When there are fewer parameters per measured variable,

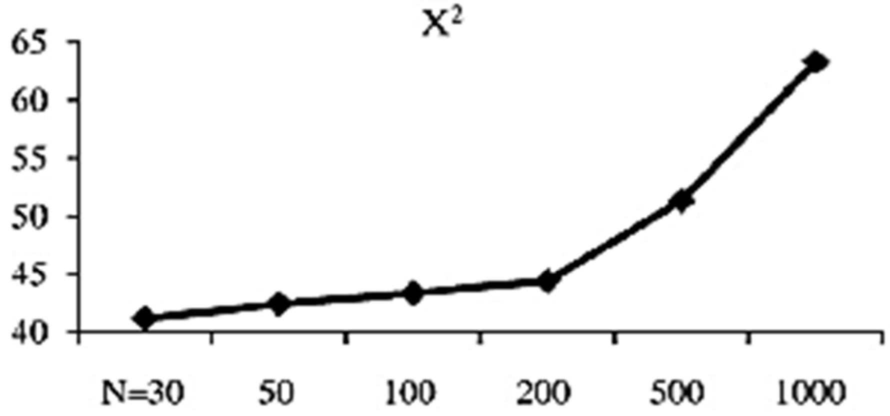
**However..., some of these rules do not hold with chi-square statistics.**

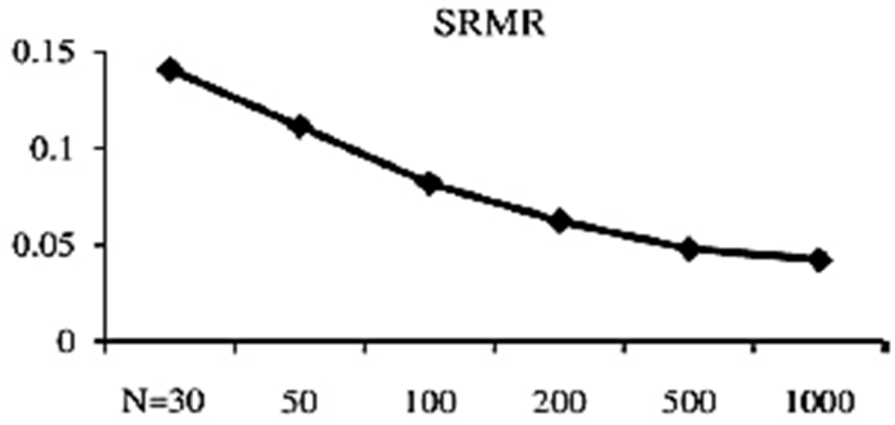
**In fact, they increase the need for sample size in order to hold chi-square rejection rates constant.**

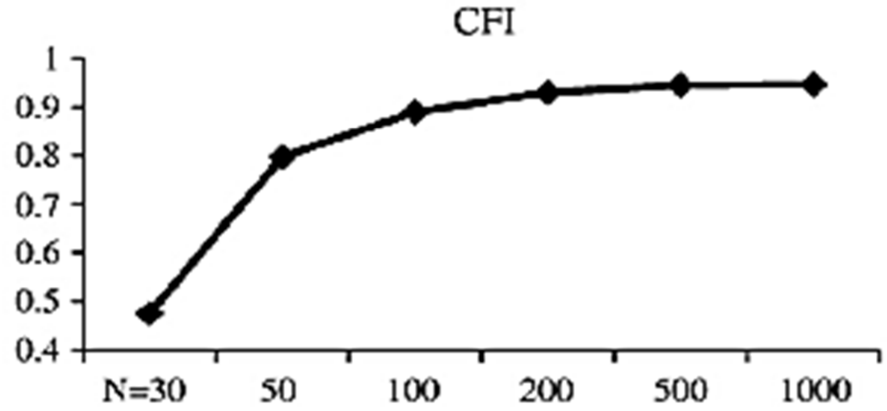
# Positive and negative factors in SEM

| | Factors that „compensate" for sample size | Factors that increase the need for sample size |
|---|---|---|
| | ✓ # of latent variables,<br>✓ # of measured variables per latent variable,<br>✓ # (fewer) of parameters per measured variable,<br>✓ saturation level | Non-normality |
| | | Non-normality |
| Chi-square statistics | | Degrees of freedom (# of latent variables, # of measured variables per latent variable, # of parameters per measured variable, saturation level, non-normality |

*Source: Marsh & Hau (1999), Hoogland (1999)*







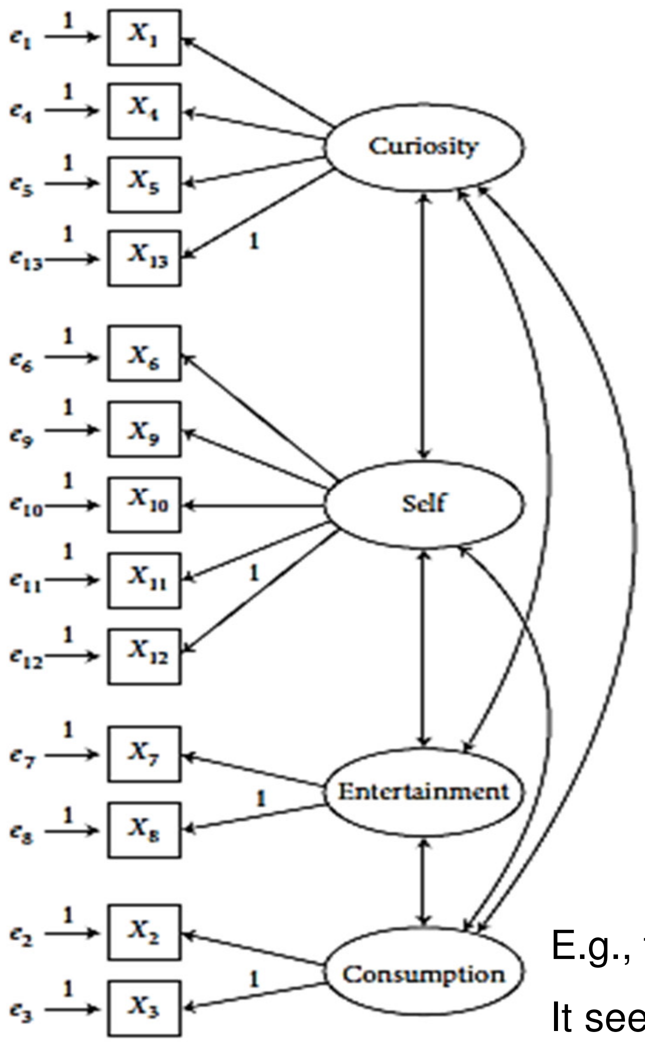
**Sample size problem?**

The above Figures illustrate Iacobucci, 2010):

- The effect of sample size on χ2 explodes for large *N* (e.g., 500 or 1000) as its corresponding *p* values decrease,
- The effect for SRMR is nearly linear – every new data point contributes to helping SRMR,
- The effect on CFI is nonlinear and data suggest that a minimal sample of 50 may be already beneficial.

# DF problem?

Scale development for hedonic-consumerism values



Source (Tarka, 2015)

(Bentler 2007, pg. 828): „in standard SEM, I am willing to believe that null hypothesis will be precisely true, but it is hard to believe to take this viewpoint in a model with large DFs. Such a model is liable always to be **misspecified**, and hence to be rejected by any „exact" test."

| | | | | | |
|---|---|---|---|---|---|
| No of indicators | 13 | 20 | 30 | 40 | 50 |
| No of factors | 4 | 6 | 9 | 12 | 16 |
| Input information | 91 | 210 | 465 | 820 | 1275 |
| Parameters | 32 | 55 | 96 | 146 | 220 |
| *DF* | 59 | 155 | 369 | 674 | 1055 |

E.g., for 369 DFs there are 369 of being correct when specifying model. It seems unlikely that any researcher would have enough knowledge to propose a model that is precisely correct in all 369 ways.

# Large data vs. test statistics - *TML*

*TML* …..problems:

- ***TML*** needs regular conditions to behave properly: *normality* and sufficiently *large sample size (N)* Yuan, Yang & Jiang (2017).

With **normally distributed data** and a **correctly specified model**, *TML* approaches a chi-square distribution as the sample size *N* increases.

This process usually ends when normality is violated, resulting in high rejection rates (Bentler & Yuan, 1999 ; Fouladi, 2000; Hu, Bentler, & Kano, 1992; Nevitt & Hancock, 2004; Savalei, 2008).

# Large data vs. test statistics – *ADF*

An alternative test statistic (ADF), although does not invoke the normality assumption (Browne, 1984).

… **it needs medium to large sample sizes to get stable estimators, and unreasonably large sample sizes to make the ADF test statistic behave as a nominal chi-square variate** Yuan, Yang & Jiang (2017).

Given this, Yuan and Bentler (1997) developed a finite sample correction to the *ADF* statistic *(YB)* that permits *ADF* testing in intermediate sample.

# Large data vs. test statistics – *TRML /TAML (SB)*

**Rescaled mean statistic (*TRML*)** and **mean-and-variance adjusted statistic (*TAML*) - SB** (Satorra & Bentler 1988, 1994), were the corrections to the likelihood ratio statistic following normal-distribution-based maximum likelihood (NML).

Note that, although both statistics have been shown to work very well in practice (e.g., Hu et al., 1992; Curran et al., 1996), an unsatisfactory aspect is that **their theoretical null distributions remain generally unknown for a nonnormal data set (Yuan et al. 2018).**

# Attention on Big Data – selected test statistics in literature?

The problem of BD addressed by Yuan, Jiang & Yang (2018) – a subject of **mean statistic (*TRML*)** and **mean-and-variance adjusted statistic (*TAML*)** .

In particular they were interested: *If TRML and TAML statistics, enjoy the properties to which they are entitled as far as the Big Data are concerned?*

Yuan et al. (2018) argued that (…)

- The mean of *TRML* can be hundreds of times greater than that of the nominal chi-square distribution in standardized units when ***p* is large** **but *N/p* is relatively small, even when data are normally distributed or when the condition of asymptotic robustness holds**.
- The mean of *TRML* can be much smaller than that of the nominal chi-square distribution when **both *p* and *N* are large**, **and if the underlying population distribution has a large relative multivariate kurtosis.**
- Similarly, the *TAML* can be far away from those of its reference distribution.

# SEM and Big Data – problems?

- While it is important to have a large sample to enhance the precision of parameter estimation in SEM, it is the case that as *N* increases, chi-square „blows up". **A chi-square e.g. related with *TML* will almost always be significant (indicating a poor fit) even with only modest sample sizes** (Iacobucci, 2009).

- Demand of **computing power which increases with *N* and *p*.**

Yuan et al., (2018) argued that SEM becomes more and more difficult to replicate the values (e.g., of the *TRML* and *TAML* statistics) as *p* increases, even with high-performance computing facilities that are available. **In their study they set maximum value of *p* at 80 variables.**

- Problem with **research costs, cost-effectiveness demands on which decision has to be made.** Such research is likely to go beyond the available resources of many scholars.

# Conclusions

**It is true that in some instances "bigger is better" when it comes to sample size,** and this assumption holds particularly when **measures are not quite clean/reliable, and the structural model does not distinguish very clearly among constructs, etc.**

**…….however, if measures are of good quality, and model is not overly complex, smaller samples will suffice** (Bearden, Sharma & Teel 1982; Bollen, 1990).

For properly specified models, as *N* increases, the fit function that connects *N* to chi-square decreases correspondingly (Bollen, 1990), and hence chi-square does not increase, and does not lead to model rejection (Hayduk, 2007).

**Bentler's (2007) recommendation:**

- rejection of SEM models with N < 200 for areas where large samples are easily available,

…however….if the small *N* is not due to laziness and the science seems appropriate, he recommended consideration of small *N* model.

# Data problems in SEM

Big Data may influence SEM in two complimentary ways:

**First:** numerous observations in dataset,

**Second**: numerous indicators / latent variables as well as parameters.

## Big data – some advantage

**A full spectrum and wide-range picture within the investigated phenomena.**

BD opens up new avenues of research and makes it possible to answer questions that were previously unanswerable in science.

# Big data – some advantage

Peter Doorn, director at Data Archiving and Networked Services in the Netherlands, when questioned:

***Do you feel that research in the social sciences is taking full advantage of the opportunities that big data currently presents?***...answered in the following way:

„**No, I think that it is so far still only a very small group that is intrigued by these new possibilities, as well as the new challenges.** The majority, however, are not. We can only speculate why. Perhaps it's because **their research questions are more traditional ones that they can solve with just a small data set......**

………Personally, I think there needs to be more demonstrator projects, which can serve as examples to the rest of the scientific community of what can be done with big data. The more projects that are carried out, the more others will see the enormous advances that are being made."

# Questions?

- Will some of the statistical (e.g., multivariate) approaches need to be re-imagined?
- Will some of the research practices in social, psychological, behavioral and other areas be discarded in the presence of new data-rich environment?
- **What is the real extent of BD influence on Structural Equation Modeling (SEM)?**
- **Do SEM researchers gain access to currently unknown features of the world by testing their models when more data appear?**

Thank you