IRT – The way forward?

Department of Public Health
University of Aberdeen
UK

Unit of Applied Clinical Research
NTNU
Norway

Peter Fayers

History

- IRT
  - Binary outcomes.
  - Estimate probabilities, not means.
  - Slow to be adopted – lack of user-friendly software.
  - Used in medical research in late 1990's
  - ……..?
- Logistic regression
  - Binary outcomes.
  - Estimate probabilities, not means.
  - Slow to be adopted – lack of user-friendly software.
  - Used in medical research 1970's.
  - First became popular in 1980's after facilities incorporated in standard packages.

Current software

- **IRT**
  - **Not in standard packages.**
  - **Must purchase standalone software.**
  - **Different programs produce different results.**
  - **Different programs fit different models**
  - **Unfriendly interfaces.**
  - **Some programs require user to specify and experiment with, e.g., “quadrature points” and “acceleration methods”, etc.**

Using IRT software is more of an art than a science.

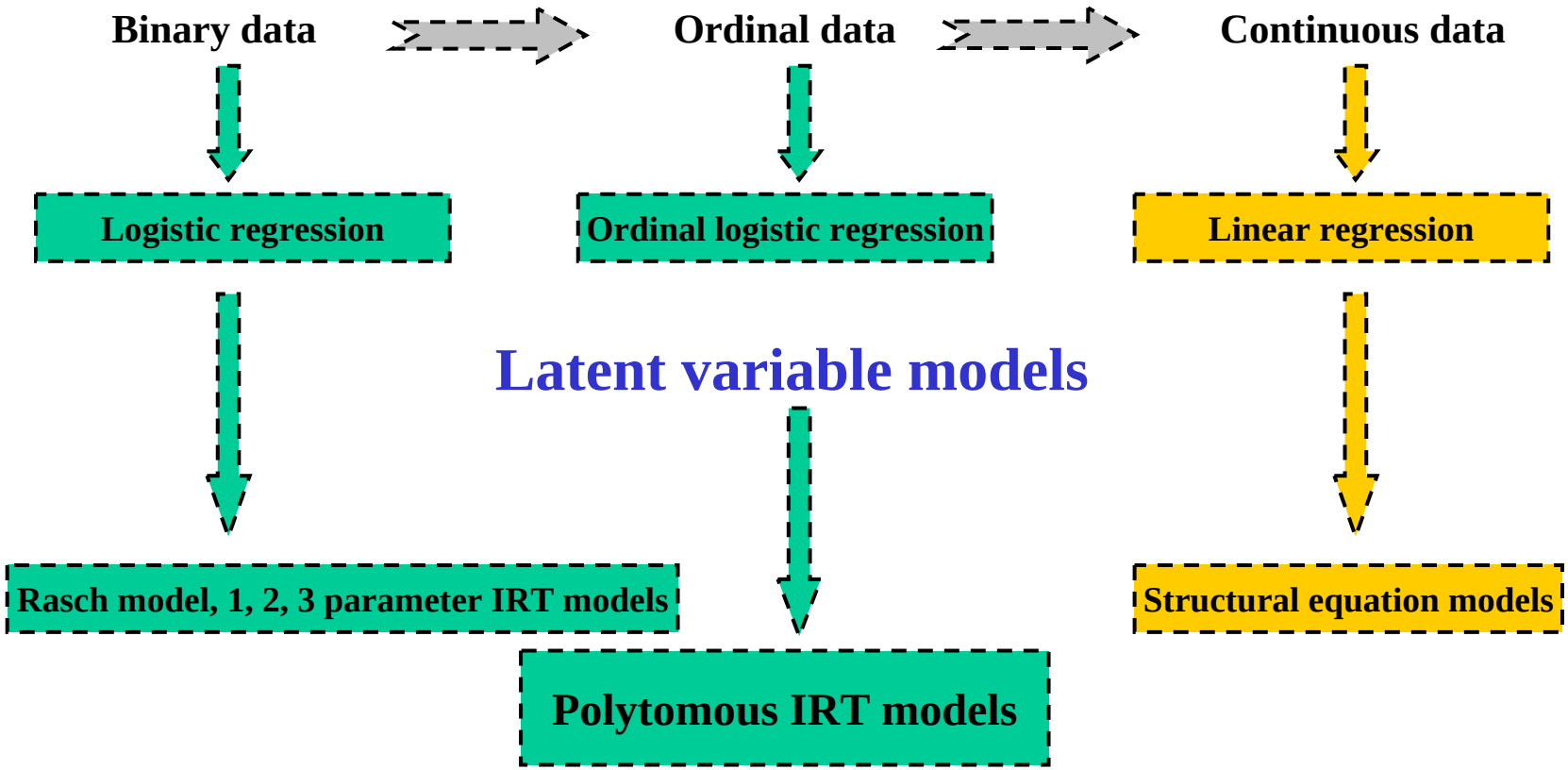
Current software – the future

- **Need for software that is:**
  - **Widely avialable**
  - **Reliable.**
  - **Easy to use.**
  - **Can fit the range of models used in PRO research.**
  - **Extensive graphical displays.**
  - **Integrated with, or having compatible data structures and interfaces with, a major statistical package.**

Generalized models

Models for observable outcomes

Probablistic models



Latent variable models

Generalized models

- **Ordinal logistic regression:**
  - **Is difficult to explain and interpret.**
  - **Choice of models is more complex.**
  - **Choice of appropriate model can be critical.**
  - **Needs larger sample size.**

- **Polytomous IRT models:**
  - **Difficult to explain and interpret.**
  - **Try one (e.g. GRM) and if it does not fit try another (e.g. GPCM) … !**
  - **Can find that different models "fit best" for different samples, even with an identical PRO.**

Need for research into linking theory with practice – e.g. can there be prior justification for selecting a particular mode, rather than trial and error.

Generalized models – the future

- **Need for research and guidelines**
  - **to explore better ways of communicating the advantages of various polytomous models.**
  - **about explaining clinical significance of coefficients (e.g. what value is conferred by having multiple response options?)**
  - **about the selection of appropriate polytomous models, and the robustness of these models.**

Assumptions of IRT models

- **Enthusiasm for IRT sometimes outweighs prudence.**
- **Few publications report the testing of assumptions – often one may suspect no tests were carried out.**
- **Example tests:**
  - **Unidimensionality**
  - **Local / conditional independence**
  - **Adequacy of the selected model**
  - **Person misfit**
  - **Item misfit**

  For polytomous models, some of the multi-response items may fit one model, other items another model.

Assumptions of IRT models

- **Little information is available about robustness of models …**
- **… or about the computer algorithms used to fit them.**
- **How sensitive are models to the various assumptions?**
- **At present, many of the available software packages only implement a few crude diagnostics.**

Assumptions of IRT models – the future

- **Need for:**
  - **Theoretical and empirical research into robustness of models applied to PRO data, and into sensitivity to assumptions.**
  - **Training courses and tutorial articles to emphasise and demonstrate the use of diagnostics.**
  - **Guidelines about usage of diagnostics.**
  - **Software that implements diagnostic tests, summary test statistics, and graphical diagnostic displays.**

Sample size

- **Early survival studies frequently far too small.**
- **Now, ethical review committees, funding bodies, regulatory authorities and medical journals *all* require evidence and reporting of pre-study sample size estimation.**
- **IRT? Few papers comment about sample size issues. Many publications about derisorily small samples.**
- **Complex polytomous models of lengthy multi-item scales may need very large samples …**

Sample size

- **Survival studies – we know how to calculate sample size and power.**
- **IRT?**
- **Largely anecdotal rules of thumb, based on experience and simulations.**
- **E.g. ½ page in Embretson & Reise (and similar coverage in other books)**
  - **E&R recommend ~500 examinees for GRM -> but "simulation studies are useful only if the data matches the simulated data".**
- **Hardly surprising that those of us doing applied research feel confused !**

Sample size – the future

- **Need for:**
  - **Further research on sample size and power.**
  - **Published guidelines.**
  - **Software to implement sample size calculations.**

- **Meanwhile,**
  - **Standard errors of estimates (or confidence intervals) are essential and should always be reported in publications.**
  - **All IRT software should provide these estimates and CIs – fortunately, many packages already do.**

IRT for PROs

- **Educational tests:**
  - ***Single dimension – e.g. math ability.***
  - ***Large number of items.***
  - ***Individual (binary) items scored "right" / "wrong."***
  - ***Infinite potential pool of items – can choose items that fit the model.***

- **PRO outcomes:**

1. Generic dimensions e.g. fatigue, pain, etc.
  - ***Typically multi-item scales.***

2. Disease-specific PROs such as symptoms.
  - ***Frequently single items.***

Educational tests vs. Clinical PROs

- **Educational tests – choose items to fit model.**
  Clinical PROs – choose model to fit the items.

- **Educational tests – lengthy tests, IRT suitable.**
  Clinical PROs – often single items less suitable for IRT.

- **Educational tests – binary items.**
  Clinical PROs – polytomous items.

- **Educational tests – assessment of ability, with dimensions / items combined objectively.**
  Clinical PROs – assessment of feelings and opinions, where only the patient can subjectively combine dimensions / items. A role for global questions?

IRT for PROs – the future

- **Need for debate and guidelines:**
- **The role of IRT for different types of PROs –**
  - **Multi-item tests for dimensions such as fatigue? IRT very useful!**
  - **Single-item or short tests for checklists of symptoms? IRT useful, but …**
- **Objective vs. subjective weightings –**
  - **IRT-based summary scores for impact of symptoms?**
  - **Global questions that allow personal weights / preferences to be subjectively applied ("Overall, how do your symptoms affect you?")**

Multidimensional IRT

- **Emotional function (EF) correlates with Physical function (PF).**
- **Therefore PF provides information about the level of EF.**
- **Imagine 2 patients, same reported level of EF, but one patient with higher reported PF.**
  MD-IRT implies that the patient with higher PF has higher "true" EF.
- **This seems to me contentious ….!**

Multidimensional IRT – the future?

- **Is there a future for Multidimensional IRT ?**

- **Probably yes but with caution?**

DIF

- IRT useful when checking for questionnaire "bias" & DIF – gender, cultural, age, etc.
- IRT (and other methods) can assess DIF.
  - Large sample ⇒ Statistically significant DIF
  - Small sample ⇒ NO statistically significant DIF
- What constitutes clinically important DIF?
- How can we communicate the impact of DIF?

DIF – the future

- **Need for research and guidelines:**
  - **Interpretation of DIF results.**
  - **Communication of results.**
  - **When does DIF matter?**
  - **How should PROs be modified if there is DIF?**
  - **Can IRT-scoring compensate for DIF, or do the questionnaires need to be changed?**

You can't make a silk purse out of a sow's ear

- IRT cannot test face or content validity.
- Qualitative methods must be rigorously applied in early phases of instrument development.
- IRT (or other psychometrics) can never salvage poor face and content validity.
- (But they can and do leave investigators blissfully unaware of the inherent design flaws in their instrument.)

Peter's law of questionnaire development

If face and content validity are high, you will end up with a good instrument.
BUT
If face and content validity are low, you will NEVER have a good instrument.

- **I have yet to see a convincing example where subsequent application of quantitative methods such as IRT ever changed this rule.**

- **Psychometric methods such as IRT merely optimize what you can build upon your initial foundations.**

Face and content validity – the future

- **Many of the widely-used PRO questionnaires have scantily reported and dubious face and content validity.**

- **IRT provides a useful tool to *supplement* but *not supplant* traditional psychometrics and qualitative methods.**

- **Guidelines on IRT in PROs should emphasize the need for prior rigorous qualitative development of instruments.**

We'll have to up your anti-depressants because, according to Dr R's latest IRT instrument, your depression is becoming much worse.

Slide prepared by Peter M. Fayers, Ph.D.

CAI

=Clinician Adaptive Interviewing

Summary

- **In many respects, IRT in PRO research is one of the most exciting developments for decades.**

- ***Like many new and novel techniques, IRT is in danger of appearing as a panacea and is in danger of enthusiastic over-use, leading to abuse and misuse.***

Recommendations

The sound application of IRT in PRO research calls for:

- **Better software –**
  - ***Easy to use.***
  - ***Standardized and readily available.***
  - ***Fitting a wide range of models.***
  - ***Good diagnostics.***
  - ***Graphics.***
  - ***Sample size.***

- **Further research into areas such as –**
  - ***Generalised models.***
  - ***Sensitivity to assumptions, robustness.***
  - ***Sample size estimation.***

Recommendations

- **Consensus view about the role of IRT for different types of PROs –**
  - ***When is the "educational model" appropriate?***
  - ***Should IRT be used for symptom checklists as well as dimensions such as EF, fatigue, pain, etc?***
  - ***What is the role for "global questions"?***

- **Published guidelines –**
  - ***When to use IRT for PROs.***
  - ***How to test and select IRT models.***
  - ***How to test assumptions, and interpret.***
  - ***Sample size.***
  - ***When does DIF matter? What to do about it?***
  - ***The important roles of qualitative methods, traditional psychometrics and IRT.***

IRT has great potential !

Slide prepared by Peter M. Fayers, Ph.D.