Comparison of item response theory and classical test theory for power/sample size calculation for questionnaire data with various degrees of variability in items' discrimination parameters

UPPSALA UNIVERSITET

Emilie Schindler, Lena E. Friberg, Mats O. Karlsson

Department of Pharmaceutical Biosciences, Uppsala University, Sweden

Objectives

To compare item response theory (IRT) and total score (TS) approaches for power/sample size calculation based on longitudinal questionnaire data for different magnitudes of variability in items' discrimination parameters. Patient-reported outcomes (PRO) are used as an example.

Background

- IRT pharmacometric modeling has been shown to increase power to detect drug effects in Alzheimer's disease [1] and multiple sclerosis [2].
- During clinical trials, PRO are increasingly collected as questionnaires containing ordered categorical items to evaluate variables not directly quantifiable, such as fatigue, health-related quality of life or pain. IRT modeling offers an alternative to classical test theory using TS to overcome challenges in PRO analysis owing to their multi-scale nature and frequent missing data.

Methods

Study design

- Parallel-group trial with a placebo and an active dose arm including 1000 simulated patients per arm and 6 occasions per subject.
- Simulated PRO data: physical well-being subscale (PWB) of Functional Assessment of Cancer Therapy-Breast (FACT-B), composed of 7 items with ordinal scores ranging from 0 to 4.

Clinical trial simulations and power calculations

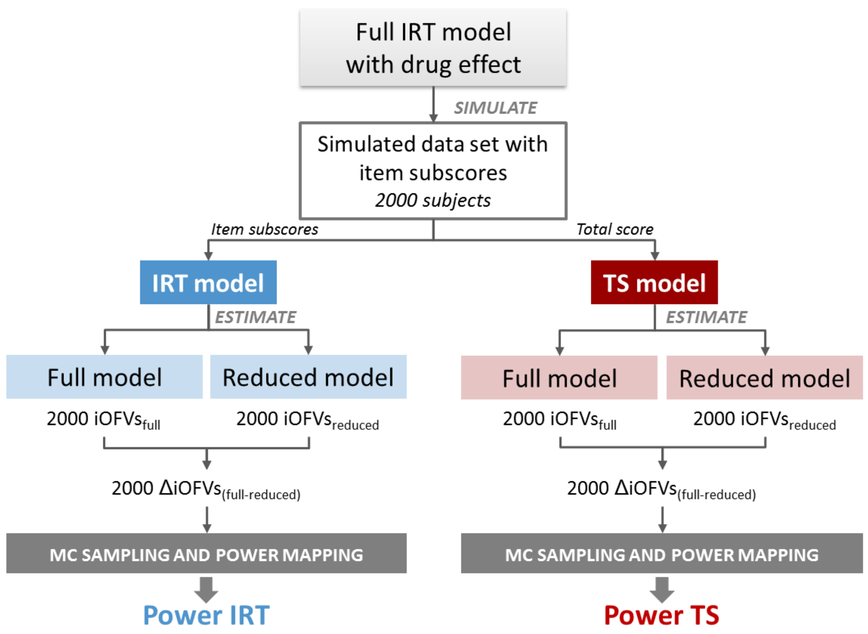

Figure 1: Clinical trial simulations and power calculation procedure using the Monte Carlo Mapped Power (MCMP) method. Full model: with drug effect; Reduced model: no drug effect; iOFV: individual objective function value; IRT: item-response theory; MC: Monte Carlo; TS: total score.

- Full IRT model with drug effect:
  - Probability of a score k to the j^{th} item for the i^{th} individual:

$$P(Y_{ij} \geq k) = \frac{e^{a_j(WB_i - b_{jk})}}{1 + e^{a_j(WB_i - b_{jk})}}$$

$$P(Y_{ij} = k) = P(Y_{ij} \geq k) - P(Y_{ij} \geq (k+1))$$

WB_i: individual latent variable for well-being; a_j: discrimination parameter for the j^{th} item; b_{jk}: difficulty parameter for score k of the j^{th} item

  - Linear effect of time on WB_i:

$$WB_i(t) = WB_{i,0} + (\theta_1 \cdot x_{arm} + \eta_i) \cdot t$$

$WB_{i,0} \sim N(0,1)$; x_{arm}: 0 for placebo arm, 1 for active dose arm; $\eta_i \sim N(0, \omega^2)$; θ_1: hypothetical drug effect (a value of 0.25/time unit was used in the simulations)

- Full TS model
  - Total scores were calculated as the sum of simulated item responses.
  - TS model: $TS_i(t) = \frac{28}{1+e^{-[Intercept+\eta_1+(SLP \cdot x_{arm}+\eta_2)\cdot t]}}$
- Simulations and power calculations:
  - Original parameters (a_j, b_{jk}) were obtained from IRT modeling of pre-treatment FACT-B PWB in metastatic breast cancer patients [3].
  - 4 scenarios: 0%, 50%, 100% and 200% of original variability in a_j across the 7 items.
  - Power calculations were performed using the Monte-Carlo Mapped Power (MCMP) method implemented in PsN software [4].

Conclusions

The value of IRT modeling over the classical test theory using total scores may increase along with the variability in items' discrimination parameters.

Results

- Item characteristic curves obtained from the parameters used for two scenarios (0% and 200% of original variability) are exemplified in Figure 2.
- IRT approach resulted in smaller sample sizes to achieve 80% power to detect a drug effect compared with TS approach in all scenarios (Table 1, Figure 3).
- The difference in required sample size between IRT and TS increased along with the variability in items' discrimination parameters.

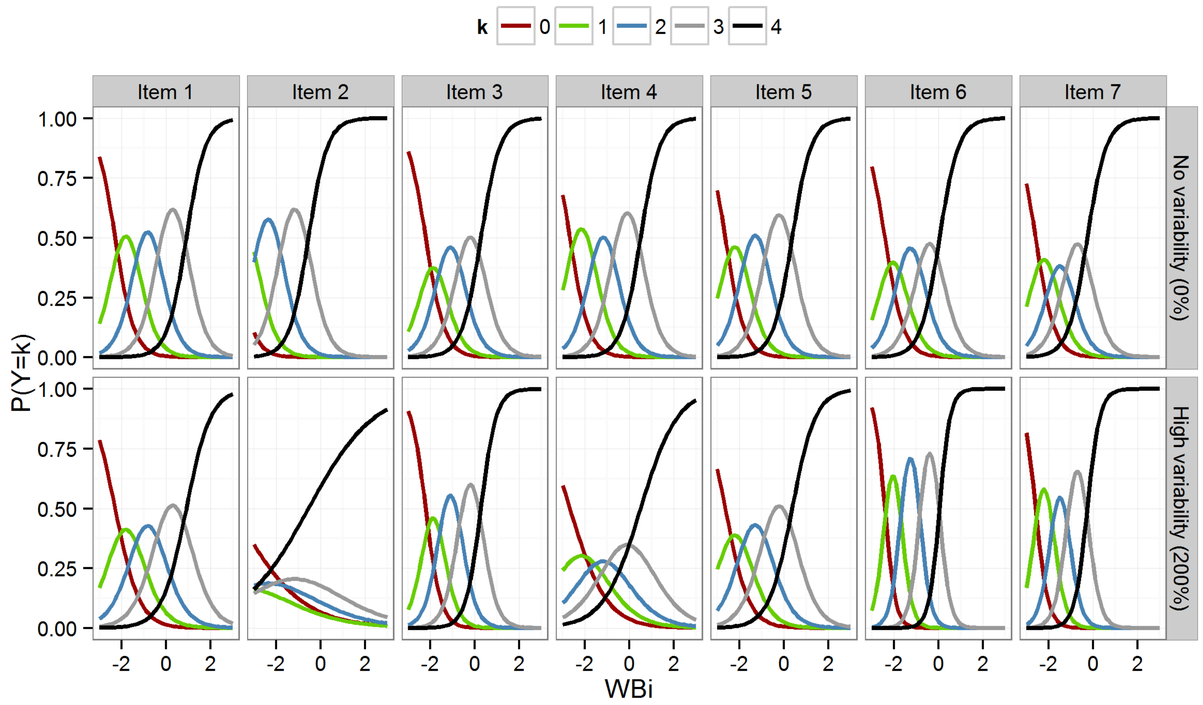

Figure 2. Item characteristic curves for 0% and 200% of original variability in items' discrimination parameters. For every item and level of variability, the probability of each score is plotted against the latent well-being WB_i.

Table 1. Total sample size required for 80% power to detect a drug effect on the latent well-being for different magnitudes of variability in discrimination parameters, using IRT and TS approches

Variability in discrimination parameters	Sample size IRT model	Sample size TS model	Difference in sample size
Original (100 %)	58	78	26% lower for IRT
No variability (0 %)	62	76	18% lower for IRT
Low variability (50 %)	64	80	20% lower for IRT
High variability (200 %)	60	100	40% lower for IRT

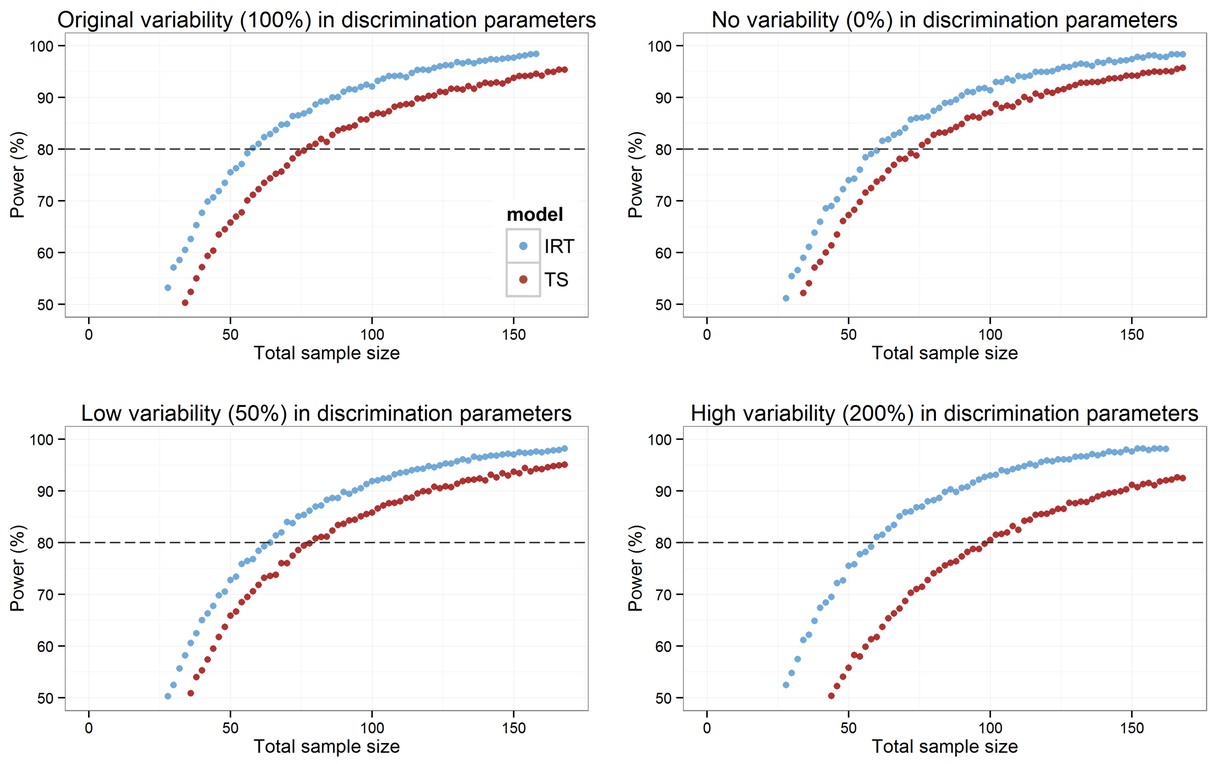

Figure 3. Power to detect a drug effect *vs* total number of subjects for different magnitudes of variability in discrimination parameters. Comparison of power curves for the item-response theory (IRT) model and total score (TS) model. The dashed line represents 80 % power.

References

[1] Ueckert S. et al. Pharm Res, 2014; 31(8): p. 2152-65
[2] Kalezic et al. PAGE, 2014; Abstr. 3262
[3] Welslau M. et al. Cancer, 2014; 120(5):642-51.
[4] Vong C. et al. AAPS J, 2012; 14(2):176-86

emilie.schindler@farmbio.uu.se