Using Optimal Test Assembly Methods to Shorten Patient Reported Outcome Measures

Daphna Harel

International Conference on Health Policy Statistics
January, 2018

Patient Reported Outcomes (PROs)

- Patient-reported outcome measures (PROs) - such as health-related quality-of-life, aspects of mental health, or functional ability - assess aspects of patients' lives that are as important to many patients as their survival.
- PROs are generally measured through multi-item questionnaires from which estimates of the underlying latent trait can be derived.
- However, patients may be asked to respond to many different scales to provide researchers and clinicians with a wide array of information regarding their experiences.
- To alleviate this burden, researchers attempt to shorten these instruments, but current methodology is under-developed and haphazard.

What do we want to do?

- To maintain the rigor of scientific research, we want three properties: reproducibility, replicability, and optimality.
- For a method to be reproducible, it must be required that two researchers should be able to apply the same procedure on the same dataset and reach the same conclusion.
- For a procedure to be replicable, it must be required that the same conclusion would be reached when applied to two datasets from the same population
- For the method to be optimal, the final selected shortened form should have the shortest possible length while still maintaining desirable attributes, and there should be no other subset of items of the same length with more desirable properties.

My procedure

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

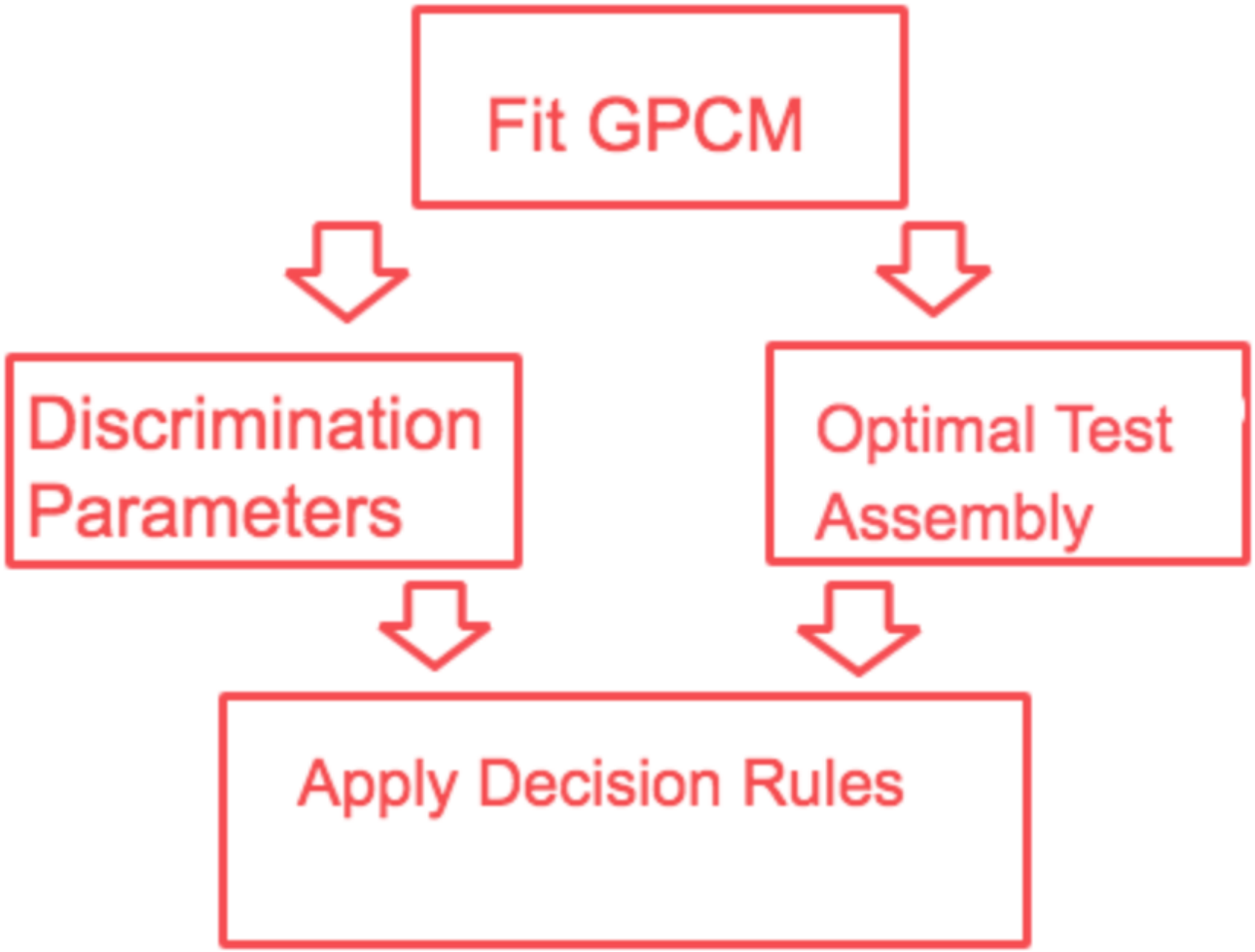
Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue



The Generalized Partial Credit Model

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

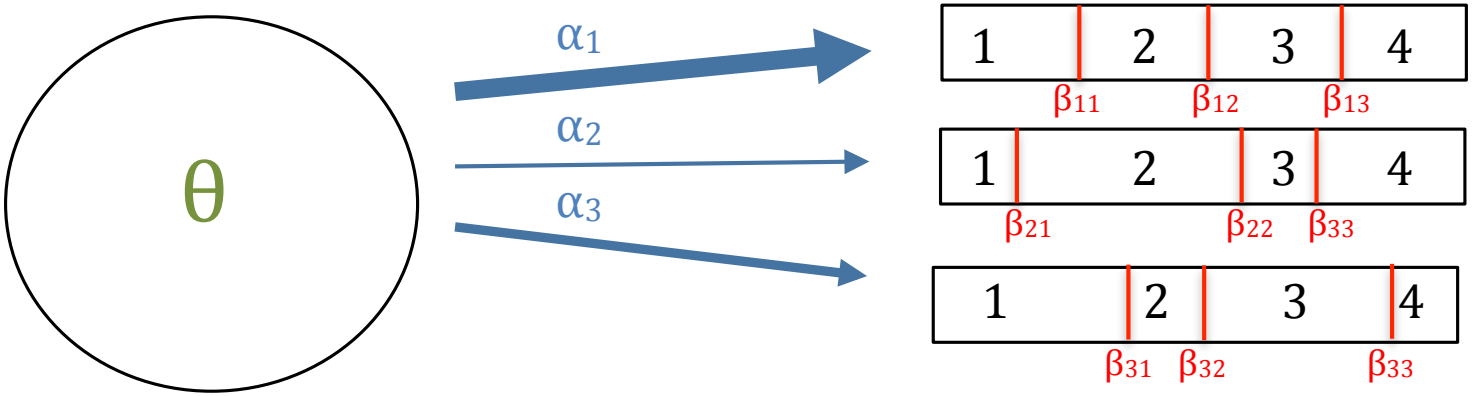
Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue



Candidates based on Discrimination Parameters

- Create a set of $J-2$ candidates from the full form by taking the items with the top j discrimination parameters for $j = 3, \ldots, J$.
- This has the effect of dropping the items with the lowest discrimination parameters successively.
- Forms of length 1 and 2 are not considered because the latent trait is not identifiable (Bollen, 1989).

Keeping it about θ

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

- Fundamentally, we have an item selection problem. But we want to select items based on our ability to estimate the latent trait.
  - Idea: Brute force! Since we've collected data on the, say, 40-item (long) version, we can try every possible subset of the instrument, get new estimates for the latent trait and see how well we've done.
  - There are $2^{40} = 1.09 \times 10^{12}$ possible forms to consider.
- Therefore we need a systematic way to think about this problem.

The Item Information Function

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

Based on Fisher's Information, the IIF represents the contribution of each item to how well the latent trait can be estimated as a function of θ.

$$
\begin{aligned}
I_j(\theta) &= \mathbb{E}[-\frac{\partial^2}{\partial\theta^2}\ell(\underset{\sim}{x};\theta)] \\
&= -\sum_{k=1}^{m_j}\frac{\partial^2}{\partial\theta^2}\ell(\underset{\sim}{x};\theta)f(\underset{\sim}{x};\theta) \\
&= \sum_{k=1}^{m_j}\alpha_j^2\left[k-\mathbb{E}(X_j|\theta)\right]^2 P(X_j=k|\theta) \\
&= \text{Var}\left(\alpha_j X_j\middle|\theta\right) = \alpha_j^2\,\text{Var}\left(X_j\middle|\theta\right)
\end{aligned}
$$

The total amount of information in the test is:

$$
I(\theta) = \sum_{j=1}^{J} I_j(\theta) = \text{Var}\left(\sum_{j=1}^{J}\alpha_j X_j\middle|\theta\right) = \sum_{j=1}^{J}\alpha_j^2\,\text{Var}\left(X_j\middle|\theta\right)
$$

Optimal Test Assembly

OTA for Shortening PROs
Daphna Harel
Introduction
Item Response Theory
Creating Candidate Forms
Discrimination Parameters
Optimal Test Assembly
Decision Rules
Simulation Study
Optimality
Replicability
Example: Fatigue

- Suppose shortened form of the PRO of length $K < J$ is to be generated.
- Define a set constraints through the creation of binary weights, $k_j \in \{0, 1\}$, that serve as indicators of whether item X_j is included in the shortened form.
- Constrain the OTA procedure to forms of length K, that is, those for which $\sum_{j=1}^{J} k_j = K$.
- If the goal of the OTA procedure is to maximize the test information of the generated form, then the specification of the problem is to maximize the linear objective function:

$$y = \sum_{j=1}^{J} I_j(\theta) k_j$$

Branch-and-bound

- To solve the constrained maximization problem, a branch-and-bound algorithm using the modified simplex method systematically searches the space of all possible shortened forms of length K to find an optimal solution that obeys the imposed set of constraints (length of form, weights are binary)
- This branch-and-bound approaches capitalizes on two facts.
  - First, adding a constraint to a maximization problem can only decrease the objective function, not increase it.
  - Second, if a solution is infeasible for a given set of constraints, meaning it does not meet the specified constraints, then it will remain infeasible if further constraints are added.

Branch-and-bound

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

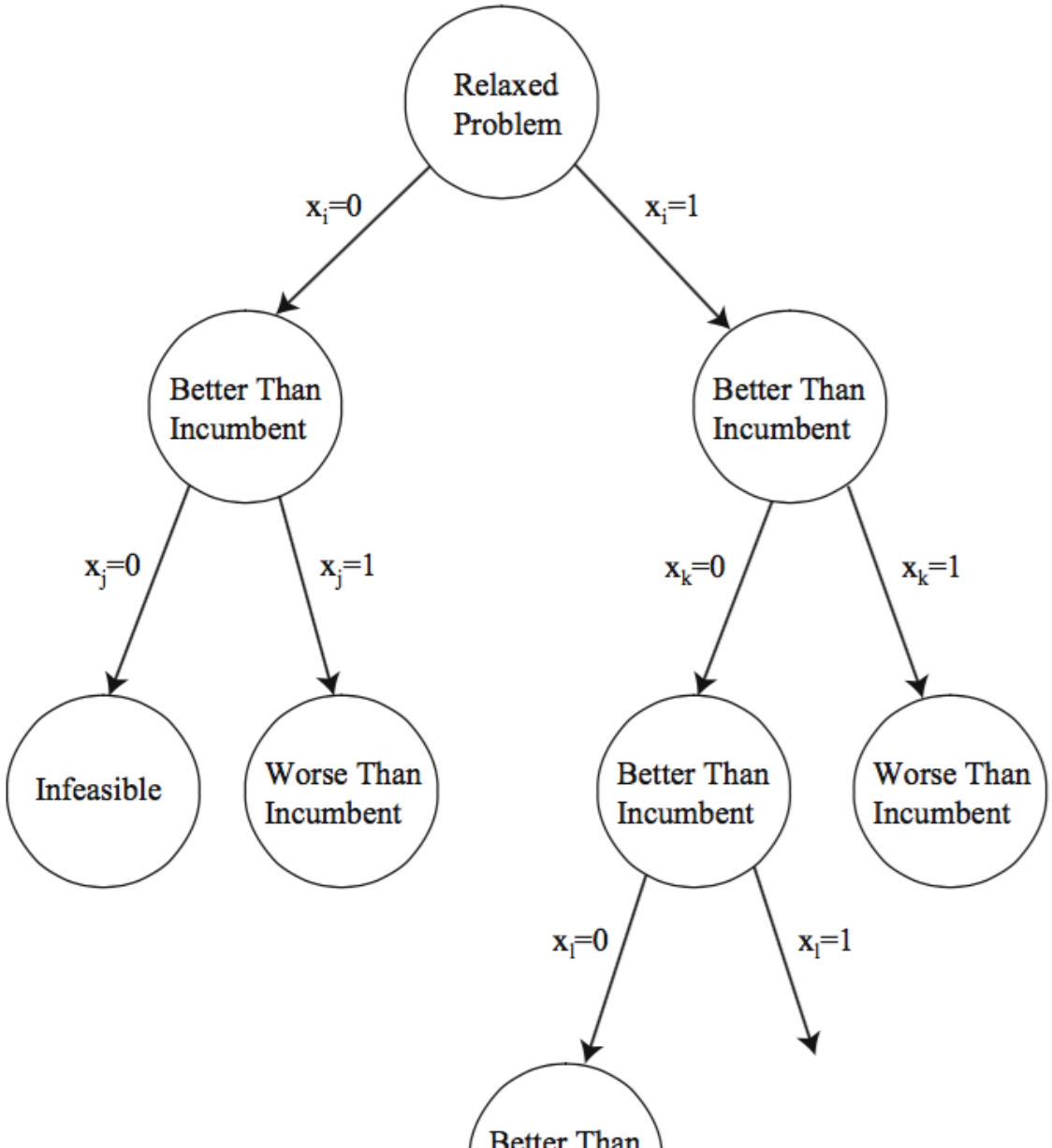
Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue



Using OTA to generate candidates

- To generate a set of candidate shortened forms of the PRO, the OTA procedure is run $J - 2$ times, constructing optimally generated, with respect to test information, shortened versions of lengths $3, \ldots, J$.
- Thus, each candidate shortened form represents the best set of items of that length to maximize the TIF.
- Now, we have two sets of candidate shortened forms, one generated through discrimination parameters, the other generated through OTA. For each, we need to select a final shortened form that is as short as possible while maintaining desirable properties.

Decision Rules: Categories

OTA for Shortening PROs
Daphna Harel
Introduction
Item Response Theory
Creating Candidate Forms
Discrimination Parameters
Optimal Test Assembly
Decision Rules
Simulation Study
Optimality
Replicability
Example: Fatigue

- Once the procedure to generate a set of candidate shortened forms of the PRO is run, it is necessary to choose which of candidate forms maintains desirable properties while still minimizing the total length of the form.
- However, there is no obvious threshold at which one would conclude that a shortened form has adequate information.
- Therefore, the properties of each candidate shortened form must be assessed in order to find a balance between shortening the scale and retaining its measurement ability.
- These properties can be formalized into rules that fall into three categories: reliability, concurrent validity, and convergent validity.

Decision Rules

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

Decision rules for reliability and concurrent validity can be operationalized as:

> ***Rule 1:*** *The candidate form is considered acceptable if it maintains 95% of the Cronbach's alpha of the full-length form.*
> ***Rule 2:*** *The candidate form is considered acceptable if the correlation between its factor scores and those from the full-length form is at least 0.95.*
> ***Rule 3:*** *The candidate form is considered acceptable if the correlation between its summed scores and those from the full-length form is at least 0.95.*

Convergent validity

Equivalence testing, which has origins in clinical trials, is used to test whether the difference between two association measures are within a pre-specified range.

$$H_0 : |\rho_1 - \rho_2| \geq \delta$$
$$H_a : |\rho_1 - \rho_2| < \delta$$

Therefore, rejecting the null hypothesis implies equivalence between the two correlations.

> ***Rule 4:*** *The candidate form is considered acceptable if the convergent validity correlation on its factor scores is equivalent, within a tolerance of* δ*, to that of the full-length form.*
>
> ***Rule 5:*** *The candidate form is considered acceptable if the convergent validity correlation on its summed scores is equivalent, within a tolerance of* δ*, to that of the full-length form.*

Simulation parameters

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

- $J = 10, 20$, $n = 250, 500, 1000, 5000$, and 500 iterations
- Thresholds either *grid* or *grouped*
- Discrimination parameters vary in homogeneity: low (0.5, 1.5), medium (0.75, 1.25), or high (0.9, 1.1)
- To assess convergent validity, a continuous measure was simulated with true correlation of $\rho = 0.7$ with the summed scores of the full-length form was generated.
- Multiple hypotheses tests for the equivalency analysis within each iteration were corrected via Benjamini-Hochberg.

Optimality

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

- Length of the final selected shortened form compared pairwise under both methods.
- Grid: Comparable performance, under both $J = 10$ and $J = 20$.
- Grouped: With $J = 10$, OTA beat discrimination parameters in 20 - 40% of simulations. With $J = 20$, OTA beat discrimination parameters in 50 - 75% of simulations

Replicability

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

- To assess replicability, Fleiss's kappa statistic was calculated to assess the level of agreement on which items were included in the selected forms across the 500 simulation conditions.
- Fleiss's kappa statistic ranges from 0 to 1, with higher values indication higher levels of agreement.
- Therefore, in this case, values near one indicate that, when considering repeated datasets from the same data-generating process, the procedure in question selected the same items to create the shortened versions.

Replicability

- When $J = 10$, the OTA procedure resulted in higher kappa values than selection based on discrimination parameters in 20 of the 24 simulation conditions, reaching moderate-to-high agreement.
- For $J = 20$, the discrimination parameter procedure resulted in higher kappa values in 16 of the 24 simulation conditions, with the largest differences observed under the low and moderate homogeneity grouped conditions.
- However, even when the discrimination parameter method had higher replicability than the OTA method, the values of the kappa statistic were still moderate-to-high for the OTA method, particularly as the sample size of respondents increased.

Three criteria to judge OTA upon

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

- Reproducibility - Yes, by default.
- Optimality - No worse, and in many cases much better
- Replicability - Only marginally worse sometimes, but still objectively reasonable.

Fatigue and Systemic Sclerosis

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

- Systemic sclerosis is an autoimmune disease characterized by microvascular disease, disturbance in fibroblast function and immune system activation, culminating in fibrosis of skin and internal organs
- Patients with systemic sclerosis have substantially reduced health-related quality of life, due in part to high levels of fatigue.
- The 13-item Functional Assessment of Chronic Illness Therapy (FACIT) scale has been validated to measure fatigue across chronic disease groups, including systemic sclerosis

Alternative: SF-36 Vitality Scale

OTA for Shortening PROs
Daphna Harel
Introduction
Item Response Theory
Creating Candidate Forms
Discrimination Parameters
Optimal Test Assembly
Decision Rules
Simulation Study
Optimality
Replicability
Example: Fatigue

- The 4-item Short Form-36 (SF-36) Vitality subscale provides an alternative to the FACIT for measuring fatigue.
- While the SF-36 Vitality subscale performs well in patients with lower levels of fatigue, it is known to not discriminate well among patients with moderate to high levels of fatigue (Harel, et. al 2012).
- This drawback of the SF-36 Vitality subscale may be due, in part, to its short length.
- Therefore, it is of interest to see whether there is a shortened version of the FACIT scale that performs equivalently to the full-length version.

Shortening!

OTA for Shortening PROs
Daphna Harel
Introduction
Item Response Theory
Creating Candidate Forms
Discrimination Parameters
Optimal Test Assembly
Decision Rules
Simulation Study
Optimality
Replicability
Example: Fatigue

- Based on a sample of 542 patients with SSc, candidate shortened forms were generated through both the OTA and discrimination parameter procedures, and the five decision rules were applied.
- To assess convergent validity, the correlation between the FACIT and the SF-36 Vitality subscale was required to be within 0.05 of the original correlation of 0.807 (95% CI, 0.775, 0.834).
- The two procedures selected the same shortened form, consisting of five items: 1, 3, 4, 5, and 6.

Is it reasonable?

OTA for Shortening PROs

Daphna Harel

Introduction

Item Response Theory

Creating Candidate Forms

Discrimination Parameters

Optimal Test Assembly

Decision Rules

Simulation Study

Optimality

Replicability

Example: Fatigue

- The type of data that is needed to do this is already usually collected when validating an original scale, so why not throw this in as the last step?
- If the convergent validity measure is dichotomous (e.g. doctor diagnoses of major depression, etc), then Ishihara et. al (in preparation) shows how you can still run the equivalency analysis.
- So far I've done this on four scales: Cochin Hand Function Scale (18 items to 6 items), Patient Health Questionnaire (9 items to 4 items), FACIT (13 items to 5 items), the Social Appearance Anxiety Scale (16 items to 5 items).
- If you have a scale you need shortened with this type of data, let's chat (daphna.harel@nyu.edu)