Detecting misfitting response patters in educational testing.

An empirical application

Rob R. Meijer, Jorge N. Tendeiro

05 July 2014 / 9th ITC Conference

university of
groningen

Overview

1. Motivation

2. Person fit analysis

3. Challenges of empirical applications

4. Data

5. Results

6. Conclusions. Limitations. Future work

Motivation

- Total scores often provide an incomplete picture of test respondents.
- Analysis of response patterns across items is desirable and recommended (ITC, 2013, p. 23).
  Advantages:
  - Better understanding of the data on the person level.
  - Clarify what unusual answering behaviors occur.
- Person-fit analysis offers various statistical approaches.

Motivation

- Idea: Compare observed with expected item score patterns.
- Expected = Based on:
  - IRT models.
  - The entire groups of respondents.
- Large differences ⟶ (potentially) misfitting or aberrant patterns.

Motivation

- A lot of overview papers and simulation studies exist.
- Empirical applications are much more sparse in published papers.
- We conducted a person-fit study based on real high-stakes educational data.
- We used existing techniques only.

Person fit analysis

- Nonparametric IRT models (NIRT; Sijtsma & Molenaar, 2002) were fitted to the data.
- Model assumptions were checked:
  - Unidimensionality.
  - Local independence.
  - Monotone IRFs.
- Useful R package: `mokken` (van der Ark, 2007, 2012).

These assumptions define the Monotone Homogeneity Model (MHM; Mokken, 1971).

Person fit analysis

- We mostly used group-based person-fit indices.
- The choice of indices was based on prior studies (e.g., Karabatsos, 2003; Meijer & Sijtsma, 2001; Tendeiro & Meijer, 2013).
- Some indices used:
  - C^* (Harnisch & Linn, 1981).
  - H^T (Sijtsma, 1986; Sijtsma & Meijer, 1992).
  - $U3$ (van der Flier, 1982).
- Useful `R` package: `PerFit` (Tendeiro, 2014).

Challenges of empirical applications

Some challenges:

1. Consider model fit.
2. Choose most adequate person-fit indices.
3. Set up reasonable cutoff scores.
4. Perform a posterior “qualitative explanation step” (Rupp, 2013).

We addressed the first three challenges in our study.
The 4th challenge was unfeasible.

Data

- Two subscales of a large-scale high-stakes educational test.
  - Section One: 23 (set-based) items.
  - Section Two: 25 items.

- All items have five response alternatives.

- $N = 4,000$ respondents.

  Perfect response vectors were removed from each Section.
  Final sample sizes:
  - Section One: $N = 3,955$.
  - Section Two: $N = 3,981$.

- Factors taken into account:
  - Gender.
  - Ratial/ethnic subgroups.

Results – Model fit

Some NIRT model-checks for both subscales:

- All inter-item covariances were positive.
  (Necessary condition; Sijtsma & Molenaar, 2002.)
- All scalability coefficients between 0 and 1.
  (Necessary condition; Sijtsma & Molenaar, 2002.)
- Monotonicity: No severe violations were found.

Results – Model fit

- Unidimensionality: We looked at
  - DETECT D (Kim, 1994; Stout et al., 1996; Zhang & Stout, 1999).
  - Scalability H (Sijtsma & Molenaar, 2002).

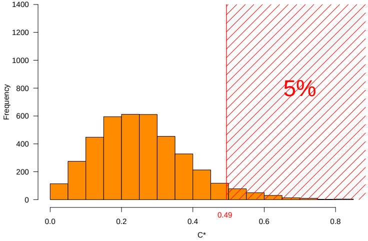
Section One	Section Two
$D = .60^a$	$D = .21^a$
$H = .20^b$	$H = .18^b$

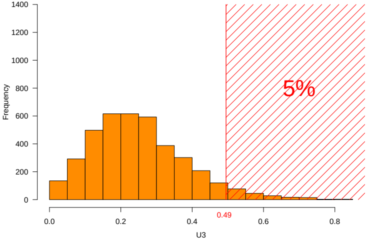
[a] Between .1 (essential UD) and 1 (MD); Stout (1990).
[b] Below the usual threshold $c = .3$.

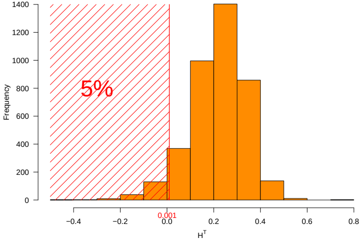
Some comments:

- Passage-based item sets might explain the dimensionality problem in Section One (not ideal).
- Item discrimination is moderate — typical of cognitive data.

Results – Person fit results (Section One)







83% of the extreme response patterns were jointly flagged by the three indices.

(Section Two: 82%.)

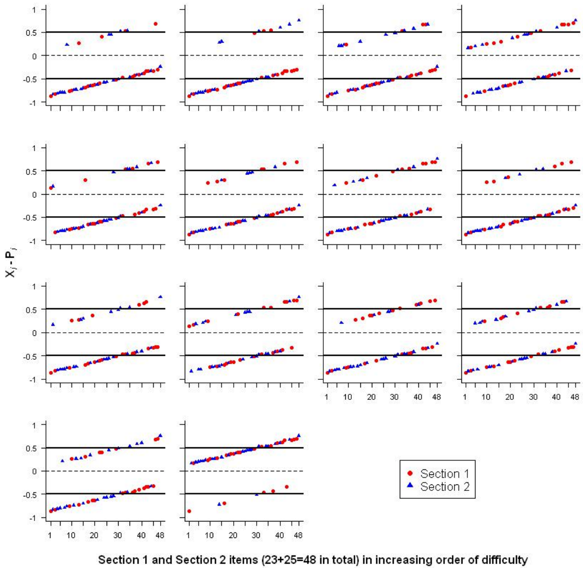
Results – Background variables

- Gender: No differences.
- First time/Retaking test: No differences.
- Ratial/Ethnic subgroups:

  One subgroup performed consistently worse on the test. It was later found that about 75% of the respondents in this groups were non-native English speakers.

Without further information, we speculate that test performance was affected by English language deficiencies.

Results – Extreme item patterns



- Many large negative residuals (i.e., incorrect answer to easy items).
- Not so many large positive residuals (i.e., correct answer to difficult items).
- Guessing may have played a role for most of these respondents.

Results – Total scores

- Total scores of flagged respondents are very close to the sample's total score mean.

- Person-fit inspections do provide added information.

Conclusions

- Inspecting item patterns provides valuable information concerning responding behavior (over and above total scores).
- Respondents with unusual response pattern were identified, interpretation of results was attempted.

Limitations

- We were unable to perform a “posterior qualitative explanation step” (Rupp, 2013).
- This is especially difficult in a high-stakes educational context.
- Other settings are more suitable for this (e.g., longitudinal settings in both educational and clinical environments).

Future work

- Set up a study which allows following up several classes of students thorugh an entire academic year.
- Conduct follow-up inspections.
  Goal: Enhance interpretation, help profiling students, provide feedback to both lecturers and students.

Thank you

Questions?