A pseudo power analysis for CTT item analysis

Alan D. Mead

Talent Algorithms Inc.

Psychometrics requires “big data”

- Samples for IRT modeling should be large
  - N >> 100
  - Although maybe N=50 for Rasch (because it only models item difficulty)
- Does CTT have lower requirements?
  - CTT difficulty SE = $\sqrt{\frac{p(1-p)}{N}}$, p=proportion correct, N = sample size
  - CTT ITC SE = $\sqrt{\frac{1-\rho^2}{N-2}} \approx \frac{1}{\sqrt{N}}$

No power analysis?

- Power analysis is the standard for determining sample size requirements
- No power analysis (that I'm aware of) for any psychometric procedure
- One reason (and this isn't obvious to everyone) is that power is only defined in terms of a hypothesis test
  - It's the "power" of correctly rejecting H0 (when H0 is false)

A “pseudo” power analysis

- Purpose: Detect poor-quality items in small samples
  - Calculate corrected item-total correlation (CITC)
  - Flag item as poor quality if CITC < $CITC_{CRITICAL}$
- $CITC_{CRITICAL}$ might be 0.0 or 0.10

A “pseudo” power analysis (cont.)

- Power = likelihood of correctly flagging poor-quality items
- Type I error = flagging high-quality item
- Type II error = not flagging low-quality item
- Heuristic
  - “Pseudo” because no distributional assumptions are being made
  - However, this method of flagging does match common practice

Power

- Ideally higher than 0.80
- Power is increased by:
  - Large samples
  - Large effect sizes
  - Choice of cut-score
- Effect size (here) is the degree of difference between population CITC of “good” and “bad” items

Simulation

- Factors
  - Number of poor quality items: 2, 5
  - Sample sizes: N = 25, 50, 75, 100
  - $CITC_{CRITICAL}$ = 0.0, 0.10
- IRT parameters:

	Good Items	Poor Items
IRT Slope	0.80	-0.50
IRT Difficulty	~ N(0,1)	~ N(0,1)
IRT Guessing	0.25	0.25

- Exam length = 25 items

SIMULATION RESULTS

Bad slope = -0.50, $CITC_{CRITICAL}$ = 0.0

	8% Flawed Items		20% Flawed Items	
N	Power	Type I	Power	Type I
25	0.97	0.17	0.95	0.21
50	1.00	0.10	0.98	0.14
75	1.00	0.06	1.00	0.11
100	1.00	0.04	1.00	0.08

- Power is excellent, even for N=50
- Type I errors are uncontrolled and excessive
- Additional flawed items degrade performance

Bad slope = 0.0, $CITC_{CRITICAL}$ = 0.10

	8% Flawed Items		20% Flawed Items	
N	Power	Type I	Power	Type I
25	0.67	0.13	0.66	0.21
50	0.79	0.08	0.76	0.10
75	0.78	0.05	0.81	0.07
100	0.81	0.05	0.86	0.05

- Power is fairly adequate for N>=50
- Type I errors are uncontrolled and large
- Additional flawed items degrade performance

Summary

- Addressed sample size requirements for detecting poor-quality items
- Framed common item analysis flagging practice as a (heuristic) hypothesis test
  - With “pseudo” Power
  - And “pseudo” Type I errors

Results Summary

- Power was good; Type I errors bad
  - Power high even in tiny samples of N=25
  - Type I errors were uncontrolled and often excessive
- Detecting items with neg. slope (-0.5) easier than items with zero slope
- Additional flawed items degraded results

Conclusions

- Surprisingly good power to detect profoundly flawed items in tiny samples
  - More modest power for moderately bad items
- Type I errors may not always be a problem
  - Labeling “bad” items as “good” might be worse
  - Possible to gather more data later

Limitations

- This “test” is very much heuristic
  - Should work out some kind of distributional test
- Only addresses one, narrow issue in CTT IA
  - DIF, Equating, etc. may not work well in small samples
- We lack a good effect size measure, but effect size had enormous impact on power
- In practice, items vary in discrimination (unless you’re a Raschian)
- No attempt to distinguish “so-so” from “good” items

Next Steps

- Restrain the Type I error rate
  - Set threshold using simulation of H0 distribution (to control Type I rates)?
  - May require assumptions about population CITC's

Thanks! Questions?

amead@alanmead.org