The perils of designing questionnaires
A critique of Finstad (2010)

Paul Cairns
Background

- I like statistics
- Frustrated by bad stats
- But it’s not just about the stats…
Finstad (2010)

- Usability Metric for User Experience (UMUX)
- Special issue of IwC on modelling UX
- Contrast with SUS
  - Shorter
  - Better language
  - 7 point not 5 point Likert scales
UMUX development

- **Pilot study**
  - 12 items split across two versions
  - 42 participants doing DB task
  - 3 items + overall ease of use item

- **Survey study**
  - Two systems: hi vs lo usability
  - 273 & 285 participants
  - UMUX and SUS

10th March, 2011
Paul Cairns, Mwahahaha!
UMUX

1. [This system’s] capabilities meet my requirements.
2. Using [this system] is a frustrating experience.
3. [This system] is easy to use.
4. I have to spend too much time correcting things with [this system].
Analysis of survey

- PCA
  - SUS and UMUX both unidimensional
- Good Cronbach $\alpha$
  - UMUX 0.94, SUS 0.97
- UMUX & SUS correlate, $r = 0.96$
- T-tests showed discriminatory power
- Further item-total correlations
Alarm bells

- Usability is 3 dimensional
  - SUS is 2
  - UMUX is 1
- 4 items?!
Reliability

- Consistency (not accuracy)
- Internal (statistical)
- Test re-test
Reliability of UMUX

- Split half method
- Cronbach $\alpha$
- Good $\alpha > 0.7$
- Too high: $\alpha > 0.8$, or 0.9
- Item-total correlations of about 0.5
Survey design

- Two systems good?
- Two systems bad!
- $r=0.9$
Validity

- Tricky in absence of objective measures
  - Efficiency, effectiveness and satisfaction
- Face: high but lacks subtlety
- Concurrent: same as SUS but…
- Predictive: but not the t-tests
- Construct: discriminatory analysis
Method for validity: Pilot

- Item-total correlations
  - Tends to produce one factor
- Usually PCA/factor analysis
  - Pilot is too small
  - Split item evaluation
- More items not bigger scales

10th March, 2011

Paul Cairns, Mwahahaha!
Method for validity: Survey

- Why not a special study?
- What PCA was done?
  - Dodgy scree plot
- Dodgy design

10th March, 2011

Paul Cairns, Mwahahaha!
So what?

- Could recover the analysis
- Stats are not so bad
- Validated against SUS?
- Stats vs science
How could this happen?

- Authors aren’t experts
- Neither are reviewers
- Nor special issue editors
  - despite what they aspire to
Psychometrics for HCI?

- Psychology vs HCI
  - Internal vs interactional
- Test re-test reliability
- Appeal of physiological measures
- Can we do better?
  - Probably but how?