Item Response Theory Applications in Health

Ron D. Hays, Discussant

These comments are supported in part by the UCLA/DREW Project EXPORT, National Institutes of Health, National Center on Minority Health & Health Disparities, (P20-MD00148-01) and the UCLA Center for Health Improvement in Minority Elders / Resource Centers for Minority Aging Research, National Institutes of Health, National Institute of Aging, (AG-02-004).

Hard to find fault with this trio



Issues

Sample Size

- Samples of 500 or more have been recommended for estimating the graded response model (Reise & Yu, 1990).
- Bonnie: n = 139 per group
- Leo: n > 10,000 per group (cross-validation)

Issues continued

Differential item and scale functioning

- Bonnie: CESD mode of administration; detection of DIF

- Leo: CAHPS English vs. Spanish; Compensatory & non-compensatory; adjustment for DIFF

Mode difference greatest in middle of scale

-- extremes not well estimated

Issues continued

Implications for scale or composite creation

- Leo: items assessing *getting care quickly* (administered using never to always response scale) produced DIFF; not items assessing *getting needed care* (administered using no problem to big problem response scale)

Issues Continued

Appropriate Model

-- Number of parameters (1-PL, 2-PL, 3-PL)

-- Dimensionality (SF-36 multidimensional--Bonnie)

Assumptions of IRT Model Tested—Unidimensionality and local independence

- Leo: linear confirmatory factor analysis

- Alternatives: categorical confirmatory factor analysis, full information factor analysis, Rasch residual factor analysis

Issues Continued

Appropriate Unit of Analysis

-- Leo: evaluated person-level but CAHPS composites used for health plan comparisons.

Cross-over Design

Mail->Phone Phone->Mail

Estimate IRT scores at both time points

Analogous to same people taking CARES, EORTC, FACT, SF-36 (Chi-Hung)

CTT and IRT results should be compared

"The critical question is not whether IRT models are superior to CTT methods. Of course they are, in the same way that a modern CD player provides superior sound when compared to a 1960s LP player…The real question is, does application of IRT result in sufficient improvement in the quality of … measurement to justify the added complexity?" Reise & Henson, J Personality Assessment, in press.

Expert ratings of stigma (Bonnie)

Lack of association between expert ratings of stigma of item and mode DIFF may mean that experts are not the best source of what items are more or less socially desirable.

All CESD items are potentially subject to response bias.

Consistent evidence that mode effects are due to SDRS bias.

Qualitative research may help to decide among competing explanations.

Issues continued

Item banking (Chi-Hung)

-- Evaluation of mode effects (CAT)

-- Maintenance of the bank (public or private)

-- Capitalize on other advantages of computer administration such as:

-> Interaction with respondent

-> Immediate feedback of results

Coming June 24-25, 2004!!

Advances in Health Outcomes Measurement:

Exploring the Current State and the Future Applications of Item Response Theory, Item Banks and Computer-Adaptive Testing

Conference co-sponsored by the

National Cancer Institute

Drug Information Association

For more information about conference, contact:

Bryce Reeve, Ph.D.

Outcomes Research Branch

National Cancer Institute

Phone: 301-594-6574

E-Mail: reeveb@mail.nih.gov

My contact information

hays@rand.org

drhays@ucla.edu

rhays@ix.netcom.com