Challenges and Best Practices in Training and Monitoring Measurement of Psychopathology in Linguistically and Culturally Diverse Settings

David G. Daniel, MD, and John Bartko, PhD

ISCTM Annual Meeting, February 22-24, 2010
Washington, DC

(Abridged Version for ISCTM Web Site Released March 5, 2010)

Financial Relationships to Disclose

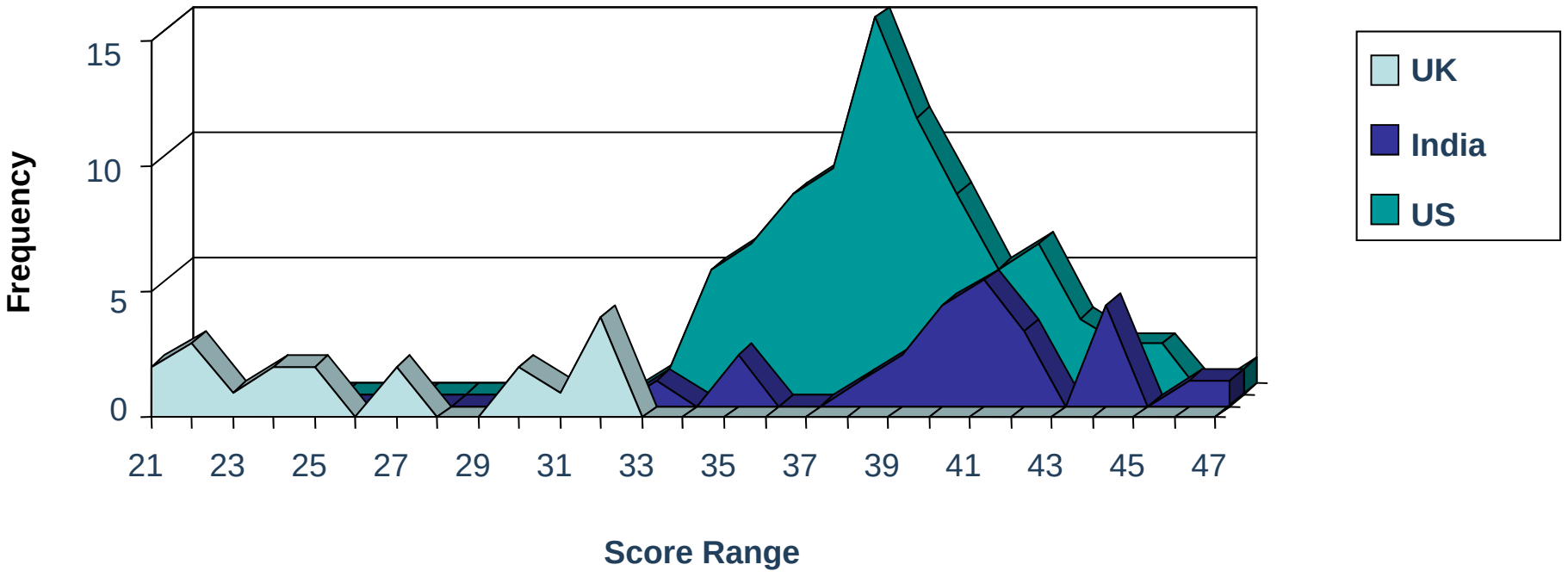
David Daniel:

- Full time employee of United BioSource Corporation
- Stockholder, United BioSource Corporation

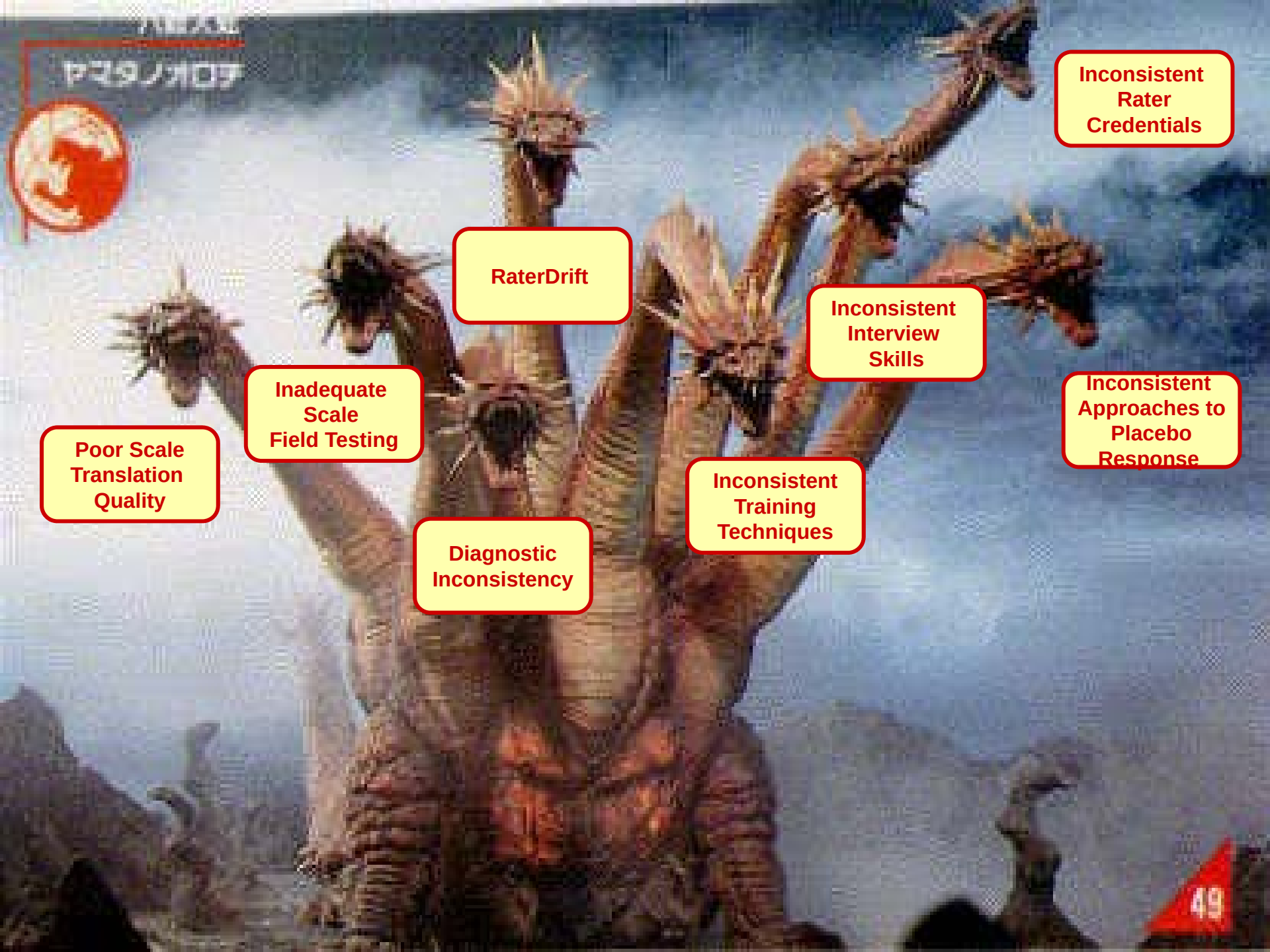
John Bartko, PhD:

- Independent Consultant

Distribution of Total YMRS Scores for U.S., European and Indian Raters



Source: "Comparison of Mania Ratings Across Three Cultures" Targum, S, Young, A, Kalali, A, Rom, D. The European College of Neuropsychopharmacology, October 2004.


Inconsistent Rater Credentials
RaterDrift
Inconsistent Interview Skills
Inadequate Scale Field Testing
Inconsistent Approaches to Placebo Response
Poor Scale Translation Quality
Inconsistent Training Techniques
Diagnostic Inconsistency
49

Modifiable Sources of Noise in Cross-Cultural Measurement of Psychopathology Measurement

- Multiple Languages and Dialects
- Inconsistent Translation Quality
- Inconsistent use of Cognitive Debriefing and Field Testing
- Variation in Credentials of Raters
- Variation in Rating Scale Experience Levels
- Inconsistent Interview Competency
- Cultural impact on Perception of Symptom Severity and Diagnosis
- Cultural Impact on Investigator Response to Training, Assessment and Remediation (Hierarchy and “Saving Face”)

What are the Best Practices For Rater Training in Multi-Cultural Contexts?

- *I. Regional vs. western centered training*
  - a. Should training and rating materials be provided in English or local language?
  - b. Should training be oriented centrally or regionally?
  - c. Should cultural effects on symptom severity be respected or harmonized with ROW?
  - II. *What are the most appropriate endpoints for training, certification and measures of inter-rate reliability?*

Variation in Translation Methodology

- ***GOLD STANDARD***
- Rigorously translate all scales, training materials and instructions
- UN style simultaneous translation at meetings
- Rigorous cognitive debriefing and local field validation of scales

- ***COMMON PRACTICES***
- Require English proficiency of rater. Scale and training materials provided in English.
- No simultaneous translation at meetings. Training in English
- Scales and instructions in English (Each rater individually translates scale materials; in the case of a PRO, rater translates for the patient)
- Informal or no cognitive debriefing. No formal field validation of scales.

Unified vs. Regional Rater Training?

- **UNIFIED TRAINING**
  - Harmonizes ratings technique across cultures
  - Reduces regional and country effects on symptom severity
  - Reduces non-specific variance and enhances inter-rater reliability across all raters
  - Allows use of same training material (often a western patient and interviewer) across cultures
  - How relevant is a western patient to a Japanese rater?
- **REGIONAL TRAINING**
  - Respects cultural variation in symptom severity
  - Training can utilize patient and interviewer of native language and culture
  - Addresses culture specific issues in rating and diagnosis
  - Enhances inter-rater reliability within region
  - Expensive and time consuming to make materials
  - Difficult to harmonize training across cultures
  - How many times should you slice the apple?

- What are appropriate endpoints for training, certification and measures of inter-rater reliability?

What Measurements Should be Standardized in Cross-cultural Training Settings?

- ***INDIVIDUAL ITEM SCORES?***
- Very sensitive to cultural variation in individual symptom severity
- Individual item variance does not necessarily affect statistical power of the primary outcome measure

- ***TOTAL SCORES?***
- Reduction in non-specific variance of the total score is relevant to sample size estimates
- Primary and secondary outcome measures are usually total scores
- Relatively insensitive to country specific interpretation of individual symptom severity?

What Measurements Should be Standardized in Cross-cultural Training Settings?

- **CROSS-SECTIONAL MEASUREMENT**

Requires one interview

Highly influenced by baseline differences in perceptions of severity

Few study efficacy outcome measures address cross-sectional assessment

- **ASSESSMENT OF CHANGE**

Requires two interviews, same patient, over time

Less biased by culture?

Most study efficacy outcome measures address change from baseline

What Measurements Should be Standardized in Cross-cultural Training Settings?

- **AGREEMENT WITH PANEL "GOLD STANDARD"**

Central vs. Regional Panels?

Subject to cultural bias of panel

Subject to experience, biases and errors by panel

Central panel gold standards may be adjusted to address regional cultural influences on scoring

- **CONCORDANCE AMONG RATERS**

Global vs. Regional?

Subject to cultural bias of the plurality

Subject to errors by raters

Directly impacts measures of inter-rater reliability

Train to Rate Videotaped or Live Subjects

- VIDEOTAPED SUBJECTS

Convenient

Efficient

Consistent training and testing across raters

Allows for calculation of inter-rater reliability

- LIVE SUBJECTS

Expensive

Allows for assessment and harmonization of interview technique and score determination

Complicated to calculate inter-rater reliability across subjects

Maintaining Rater Calibration Over Time in Cross-Cultural Settings

- Training of raters must be culturally sensitive to persist over the course of a clinical trial
- Forced shifts from entrenched cultural practices tend to regress back to habitual practices over time
- Frequent data monitoring, review of recorded patient interviews, and feedback reinforce and maintain cross-cultural harmonization of ratings and interview practices over time
- Native language experts synchronized to global practices may be the best sentinels for ongoing feedback to raters
- Frequent refresher or recertification procedures may reduce rater drift

What Symptom Measurements are Most Challenging to Standardize Cross-Culturally?

- Socially "sensitive" behaviors involving sexuality, aggression and behavior toward authority?
- Negative (compared to positive) symptoms of schizophrenia?
- Scale items evaluated solely based on observations during the interview (vs. items that require data from an informant)?

Anecdotal Observations on Cross-Cultural Rater Training

- Co-training among experts from each culture provides linkage to global rating practices while incorporating local tradition and respecting local sensitivities
- In addition to regional and country differences, the cultural distinctions within regions and countries may be very large and significantly affect rating practices
- The manner in which feedback is given affects its acceptability. For example, failure of an individual to pass certification may symbolize loss of face in some cultures.
- Hybrid meetings utilizing local experts and regional breakouts combined with plenary sessions are a desirable way to balance local practices and global harmonization

Anecdotal Observations on Cross-cultural Rater Training: Measuring Rater Proficiency and Agreement

- The metric for evaluating skills and agreement among raters may have a significant effect on certification rates and inter-rater reliability
- Western centered panels may produce skewed measures of proficiency in international trials
- Culturally neutral measures of concordance, such as agreement with the mode, may be useful in int'l trials
- Total scores and change in total scores may be more relevant and less sensitive to cultural influences than individual items
-

Anecdotal Observations on Cross-cultural Rater Training: What Symptoms are Hardest to Standardize?

- Measurement of socially sensitive symptoms involving sexuality, aggression and attitudes toward authority may be difficult to standardize
- Scale items based solely on interview observations may be harder to standardize than those utilizing both interview and informant data