# Practicality of Person-Fit Statistics in Detecting Unusual Testing Behavior

Rory McCorkle – Project Management Institute (PMI)
Shungwon Ro and Larissa Smith – Prometric



ATP Innovations in Testing 2010
Orlando, Florida

# Person-Fit Statistics: Background

- Attractive tools to detect “unusual testing behavior” or “person-misfit”
- Attentions to the analyses of person-misfit or person-fit using these tools are relatively new
- 1970’s: item response theory (IRT) models first used to identify the cause of misfit, based on person-fit methods and non-model based methods
- 2000’s: increased interest in test security with advent of computer based testing results in new focus on person-fit

# Person-Fit Statistics: Background

- Focus of research studies since 1970's has been:
  1) Identifying causes of misfit
  2) Evaluation with respect to relative detection power
  3) Distributional characteristics of person-fit statistics
  4) Major factors affecting performance of person-fit statistics
  5) Relationship between misfit and the validity of scores

# Person-Fit Statistics: Background

- Five groups of sources of misfit that potentially lead to:
  - Inaccurate measurement of traits
  - Invalid score interpretation
  - Additional consequential effects, such as
    - Losses in organizational productivity by excluding individuals from opportunities
    - Failure in screening individuals

# Person-Fit Statistics: Sources

- By test-taking strategies
  - Cheating
  - Faking  good or bad (social desirability or malingering)
  - Guessing by less able persons (lucky guessing or random guessing
- By examinee behavior
  - Answer-sheet alignment errors
  - Application of erroneous rules or misconceptions
  - Carelessness by more able persons
  - Exceptionally creative or novel interpretation
  - Fumbling (sets of incorrect responses at the beginning of the test)
  - Mismarking the answer sheet
  - Plodding (half correct and the other half incorrect)
  - Sleeping (incorrect responses to easy items /correct responses to hard items)

# Person-Fit Statistics: Sources

- Sources by examinee external conditions
  - Cultural bias
  - Cultural and language discrepancy
  - Inappropriate schooling
  - Lack of language proficiency
- Sources by faultiness in the test or scoring
  - Deviation from unidimensionality
  - Item bias
  - Mismarked scoring key and input error by scorer
  - Multiple correct answers (or no answer) on a multiple-choice item
  - Poor or attractive item distracters

# Person-Fit Statistics: Sources

- Not all person-fit statistics have been known to help identify some types of misfit and interpret misfitting item response patterns
- Most person-fit statistics compare an observed response pattern against expected response pattern; expected response pattern is determined based on an IRT model or no model given a sample group

# Person-Fit Statistics: Statistic Examples

- Examples of non-model based person-fit statistics
  - Predictor of predictability (Ghiselli, 1960)
  - Weighted average index (Jacobs (1963)
  - Personal biserial correlation (Donlon and Fischer, 1968)
  - Caution index (Sato, 1975)
  - Modified caution index (Harnisch and Linn, 1981)
  - Norm conformity indices and Individual consistency index (Tatsuoka and Tatsuoka, 1983)
  - Agreement, disagreement and dependability indices (Kane and Brennan, 1980)
  - *H* (Sijtsma, 1986)
  - *U* and *ZU* (van der Flier, 1980, 1982)

# Person-Fit Statistics: Statistic Examples

- Examples of IRT-model based person-fit statistics
  - *U* and *W* – mean-squared residual- based statistics (Wright and Stone, 1979; Wright and Masters, 1982)
  - *UB* and *UW* – unweighted between and within subset statistics (Smith, 1995)
  - $L_o$ and $L_z$ - likelihood based and standardized likelihood based statistics (Levine and Rubin, 1979; Drasgow et. al. 1985)
  - *M* (Molenaar and Hoijtink, 1990)
  - *T* (Klauer, 1991)
  - $ECI_z$ (Tatsuoka, 1984)
  - *D* (Weiss, 1973)

# What makes some mathematically sophisticated IRT based person-fit statistics distinct from others based on classical methods?

- Non-model based person-fit statistics generally tend to correlate highly with number-correct scores
- Non-model based person-fit statics are sample specific – i.e., detecting misfit response patterns is affected by group characteristics of examinees
- Using an IRT model's strong assumptions with respect to a person's item responses, it is possible to show whether a response pattern is in accordance with the model. Lack of fit to a model is stated as the degree of misfit.

# How to design and use person-fit statistics for checking aberrant test results?

- A testing program manager is unsure of test results and consider possible security breach
- Is decided to conduct person-fit analyses on all candidates/examinees
- Selected scoring model is the three-parameter logistic IRT model (3PLM)
- Person-fit statistic used is $L_z$

# How to design and use person-fit statistics for checking aberrant test results?

- 3 PLM characterizes the relationship between the probability of a person's response to an item in the keyed direction and the latent trait (theta) of that person as a logistic function, called item response function;

$$P_{ij}(\theta_j) = c_i + \left\{ \frac{1 - c_i}{1 + \exp[-a_i(\theta_j - b_i)]} \right\},$$

Where

$i$ = item 1, 2, …, $n$,

$a_i$ = the item discrimination parameter

$b_i$ = the item difficulty parameter

$c_i$ = the item pseudo-guessing parameter

$\theta_j$ = the person parameter; the location of the person along the trait continuum

# How to design and use person-fit statistics for checking aberrant test results?

1. Compute the likelihood, $l_o$ : A single value of the natural log of the likelihood function at its peak for a given $\hat{\theta}$ (or $\theta$) and response vector

$$l_0 = \sum_{i=1}^{n} \left[ u_{ij} \ln P_i\left(\hat{\theta}_j\right) + (1 - u_{ij}) \ln \left\{ 1 - P_i\left(\hat{\theta}_j\right) \right\} \right],$$

where

$u_{ij}$ = a correct response by a person $j$ to an item $i$,

$P_i\left(\hat{\theta}_j\right)$ = Probability of correct response to an item $i$ given a single fixed value of $\hat{\theta}$ in a dichotomous IRT model,

$i$ = item, 1, 2, …, $n$, and

$j$ = person, 1, 2, …, $N$.

# How to design and use person-fit statistics for checking aberrant test results?

2. Compute a standardized $l_o$ (i.e., $l_z$) for a dichotomous IRT model

$$l_z = \frac{l_o - E(l_o)}{\sqrt{V(l_o)}},$$

Where $E(l_0) = \sum_{i=1}^{n}\left[P_i\left(\hat{\theta}_j\right)\ln P_i\left(\hat{\theta}_j\right) + \left\{1 - P_i\left(\hat{\theta}_j\right)\right\}\ln\left\{1 - P_i\left(\hat{\theta}_j\right)\right\}\right]$

And $V(l_0) = \sum_{i=1}^{n} P_i\left(\hat{\theta}_j\right)\left\{1 - P_i\left(\hat{\theta}_j\right)\right\}\left[\ln\left\{\frac{P_i\left(\hat{\theta}_j\right)}{\left\{1 - P_i\left(\hat{\theta}_j\right)\right\}}\right\}\right]^2$

# How to design and use person-fit statistics for checking aberrant test results?

- $L_z$ is approximately standard normally distributed.
- $L_z$ is standardized, meaning that 0.0 reflects a perfectly typical response string. 2.0 or above indicate unexpectedly good fit (overfit). -2.0 or below indicate unexpectedly poor fit (misfit)
- Each person's response pattern is then evaluated using the calculated $L_z$ value and the criteria for misfit and/or overfit

# Limits and practicality of person-fit analyses

- The objective of person-fit analyses is to detect unusual item response pattern based on a scoring model or the majority of patterns in the sample of interest.
- "Unusual" is expressed in both directions: misfit and overfit.
- Limits

  - Existing person-fit research studies have focused on developing person-fit statistics, but not a single statistic can help identify the causes of misfit listed earlier.

  - Most of them are theoretically well grounded, but practically not sound due to true trait scores dependency or sample dependency.

  - Not all person-fit statistics appear to show few false identification of misfit response patterns. Detection rates vary across statistics and are depending on types of misfit, theta value and test length.

# Limits and practicality of person-fit analyses

- Limits

  - Few research studies with the existing person-fit statistics have been done using empirical data.

  - These limits lead to practical questions remained unanswered.

- Practicality

  - Detection rates of many person-fit statistics are often based on unrealistic distributional characteristics. This makes these tools impractical. → More research is necessary.

  - Whether using person-fit statistics routinely is justifiable in a practical situation? → It depends on the context and the types and degree of misfit a researcher tries to identify in practice.

# Limits and practicality of person-fit analyses

- Practicality

  - With well targeted types and degree of misfit designed, some person-fit statistics such as *H* might be useful.

  - Some IRT model based statistics such as $L_z$ or mean square fit statistics can be used as routine checking tools in practice, but not to prove inappropriateness of the measure.