Evaluating Model Fit with IRT

American Board of Internal Medicine
Item Response Theory Course
How good is our model?

• Assessment of Model-data Fit
  – Model Assumptions
  – Model Predictions vs. observed data
  – Examples of Goodness-of-Fit
Model-data Fit

• George Box (1979):
  “All models are wrong but some are useful.”
• A “model” is something we use to approximate reality for the purpose of making predictions, explaining data, etc.
• Strictly speaking, no model will fit the data perfectly…the question is, “how much model-data misfit is too much?”
Model-data Fit

• “How much misfit is too much?”
• Generally, there are no absolute criteria for model-data fit.
• “Let your conscience be your guide”
  – Conduct a variety of analyses, consider the full set of results, and be mindful of these with regard to the intended application of the IRT model.
Model-Data Fit

- Evaluate model assumptions.
- Assess residuals and standardized residuals and examine consequences of model misfit.
Types of Evidence

• Are *model assumptions* met to a ‘reasonable degree’ by test data?
  – Determine if *model features*, like parameter invariance, are present.

• Assess the accuracy of *model predictions* versus actual data.
  – Do expected probabilities match with empirical relative frequencies?
Model Assumptions

• Are model assumptions met to a reasonable degree by test data?
  – Unidimensionality
  – Local Independence
  – Nature of the ICC
    • Minimal Guessing (for 1- & 2-PL)
    • Equal Discrimination (for 1-PL)
  – Parameter Invariance
Unidimensionality

- Dimensionality is often analyzed through a Principal Components Analysis (PCA) to determine how many “components” or “dimensions” are necessary in order to explain a significant amount of total variance.
PCA

• PCA is mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

• Analysis is performed on the correlation matrix.

• The first principal component accounts for as much variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.
PCA

• Each principal component will receive an ‘eigenvalue’ ($\lambda_j$) which indirectly represents the amount of variance that component explains.

• In IRT, we look for a “dominant” first factor in order to infer the unidimensionality of the test.
Unidimensionality

• A *scree plot* is often used to evaluate the dominance of the first factor.

• Rules of thumb for “essential unidimensionality.”
  – First eigenvalue: $\lambda_1 / n \geq 0.20$.
  – Ratio of $\lambda_1$ to $\lambda_2$ should be “large.”
  – Other eigenvalues: $\lambda_j \leq 1$;
Scree Plot for a 40-item Test

\[
\sum_{j=1}^{n} \lambda_j = n \quad \text{so if} \quad \frac{\lambda_j}{n} \approx 0.20
\]

then we say the test is "essentially unidimensional"
The first component accounts for 21% of the total variance, and it accounts for 5.63 times the amount of variance as the second component.
Local Independence

• Item responses should be independent given ability.
  – Omits or unexpected low response probabilities for the last few items may indicate speededness of the test and will violate local independence.
  – Response probabilities should not be related to group membership (DIF).
Nature of the ICC

• The probability of success should be monotonically increasing with $\theta$.

• No systematic errors of prediction should be apparent.
Nature of the ICC

• Minimal Guessing (1- or 2-PL)
  – Check performance levels of low ability examinees.
  – Conditional p-values for low ability examinees should be close to zero when fitting a 1- or 2-PL model.
Nature of the ICC

• Equal Discrimination (1-PL)
  – Essentially homogeneous distribution for the item-total correlations.
  – All $r$ values should be approximately equal when fitting a 1-PL model.
  – When examining the model-data fit, is there any pattern of errors that might be accounted for by a different slope?
Parameter Invariance

- Determine in model parameter invariance is present.
  - Invariance of Ability ($\theta$).
  - Invariance of Item parameters ($a, b, c$).
Parameter Invariance

• Invariance of ability (θ)
  – Compare examinee ability estimates based on different sets of items from the pool of items calibrated on a common scale.
  – If the model fits, different item samples should still produce close to the same ability estimates (SE will change for shorter tests, though).
Parameter Invariance

• Invariance of ability (θ)
  – Scatterplots of examinee ability based on one sample of items versus the other should be strongly linearly related.
  – Relationship won’t be perfect: sampling error does enter the picture.
  – Those estimates far from the best-fit line represent a violation of invariance.
Parameter Invariance

• Invariance of Item parameters (a,b,c)
  – Compare item statistics obtained in two or more groups (e.g., high and low performing groups; Ethnicity; Gender).
  – If the model fits, different examinee samples should still produce close to the same item parameter estimates.
Parameter Invariance

• Invariance of item parameters (a,b,c)
  – Scatterplots of b-b, a-a, c-c based on one sample of examinees versus the other should be strongly linearly related.
  – Relationship won’t be perfect: sampling error does enter the picture.
  – Those estimates far from the best-fit line represent a violation of invariance.
Check Item Invariance

• Differential Item Functioning (DIF)
  – Details on Thursday

• A test item is labeled with “DIF” when examinees with equal ability, but from different groups, have an unequal probability of answering the item correctly.

• This violates parameter invariance because items aren’t unidimensional.
DIF issues

- **Uniform DIF**: item is systematically more difficult for members of one group, even after matching examinees on ability ($\theta$).
- **Cause**: shift in b-parameter.
DIF issues

• Non-Uniform DIF: shift in item difficulty is not consistent across the ability continuum.

• Increase/decrease in P for low-ability examinees is offset by the converse for high-ability examinees.

• Cause: shift in a-parameter.
$P(u=1|\theta)$
Residual Analysis

• Assess the accuracy of **model predictions** versus actual data
  – Residual = difference between observed proportion and predicted probability:

\[
 r_j(\theta) = P_j(\theta) - E[P_j(\theta)]
\]
Residual Analysis

\[ r_j(\theta) = P_j(\theta) - E[P_j(\theta)] \]

In this notation,
\[ P_j(\theta) = \text{observed proportion correct for a given } \theta \text{ level} \]
\[ E[P_j(\theta)] = \text{expected proportion correct (i.e., probability from the IRT model)} \]
Residual Analysis

\[ r_j(\theta) = P_j(\theta) - E[P_j(\theta)] \]

White = \( P_j(\theta) \)

Black = \( E[P_j(\theta)] \)
Standardized Residuals

• Raw residuals do not take into account the error associated with the expected proportion correct, so we standardize each by dividing by its standard error:

\[
SE(E[P_j(\theta)]) = \sqrt{\frac{E[P_j(\theta)E[Q_j(\theta)]}{N(\theta)}}
\]
Standardized Residuals

\[ SR_j(\theta) = \frac{P_j(\theta) - E[P_j(\theta)]}{\sqrt{E[P_j(\theta)] E[Q_j(\theta)]}/N(\theta)} \]

SR values should be homoscedastic for each item and follow an approximately standard normal distribution across all items of the test.
Probability of Correct Response vs. Ability ($\theta$) for the 2-Parameter Logistic Model (2-PL)
SR values are homoscedastic for this item when fit by a 3-PL model, but systematic errors are present for the 1- and 2-PL models.
Test-level Fit

• Similar to the comparison done for individual items, but instead we compare Expected Proportion Correct (TCC) to observed proportion correct (raw score/N).

• SRs across the test should be homoscedastic and follow an approximate normal distribution.
Across all items, SR values are approximately normally distributed when fit by a 3-PL model, but more uniform for the 1- and 2-PL models.
Significance Testing

Q1 chi-square (Yen, 1981)

\[ Q1_j = \sum_{i=1}^{m} SR_{ij}^2 \quad Q1 \sim \chi^2 \quad df = m - p \]

\( m = \# \text{ of quadrature points} \)

\( p = \# \text{ of item parameters} \)
Chi-square test

• Standardized residuals are essentially prediction errors that have been turned into z-scores.

• The sum of squared z-scores follow a Chi-square distribution.
  – Much like Sums of Squares and variances follow a Chi-square distribution in ANOVA.
As the degrees of freedom increase, the sum of squared standardized residuals are less likely to be equal to zero.

The expected value for $\text{SUM}(SR^2)$ moves to the right as the distribution approaches normality.
Goodness-of-Fit

• This is what we call “goodness-of-fit”:
• We hope that the chi-square test will NOT be significant:
  – This indicates that the differences between observed and expected is small.
  – Significant differences would mean that observed proportions are far from what the model predicted…and that’s bad.
Significance Testing in BILOG and PARSSCALE

• The goodness of fit information contained in BILOG and PARSSCALE use the Chi-square test described in the previous slides.

– These values can be found for all items.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>INTERCEPT S.E.</th>
<th>SLOPE S.E.</th>
<th>THRESHOLD S.E.</th>
<th>LOADING S.E.</th>
<th>ASYMPTOTE S.E.</th>
<th>CHISQ</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATH01</td>
<td>1.041</td>
<td>0.651</td>
<td>-1.599</td>
<td>0.545</td>
<td>0.186</td>
<td>29.0</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>0.107*</td>
<td>0.082*</td>
<td>0.242*</td>
<td>0.069*</td>
<td>0.084*</td>
<td>(0.0007)</td>
<td></td>
</tr>
<tr>
<td>MATH02</td>
<td>2.230</td>
<td>0.600</td>
<td>-3.717</td>
<td>0.514</td>
<td>0.199</td>
<td>9.5</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>0.165*</td>
<td>0.114*</td>
<td>0.610*</td>
<td>0.098*</td>
<td>0.089*</td>
<td>(0.0920)</td>
<td></td>
</tr>
<tr>
<td>MATH03</td>
<td>0.428</td>
<td>0.693</td>
<td>-0.618</td>
<td>0.569</td>
<td>0.159</td>
<td>63.2</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>0.106*</td>
<td>0.084*</td>
<td>0.190*</td>
<td>0.069*</td>
<td>0.071*</td>
<td>(0.0000)</td>
<td></td>
</tr>
<tr>
<td>MATH04</td>
<td>-0.601</td>
<td>1.391</td>
<td>0.432</td>
<td>0.812</td>
<td>0.217</td>
<td>10.7</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>0.216*</td>
<td>0.268*</td>
<td>0.095*</td>
<td>0.156*</td>
<td>0.040*</td>
<td>(0.2204)</td>
<td></td>
</tr>
</tbody>
</table>
The problem of sample size

- Statistical tests of model-data fit present an interesting duality:
  1) Due to sensitivity to sample size, almost any departure of data from the model results in rejecting $H_0$.
  2) For small samples, model-data misfit can be overlooked, and SEs for item parameters are large.
Standardized Residuals

\[ SR_j(\theta) = \frac{P_j(\theta) - E[P_j(\theta)]}{\sqrt{E[P_j(\theta)] E[Q_j(\theta)]}/N(\theta)} \]

As \( N \) increases, \( SE \) decreases, \( SR \) increases! Often, when you have a large enough sample size to estimate parameters, model misfit is a foregone conclusion!
Conclusion

• This leads us back to the beginning… recall what George Box said!

• Generally, there are no absolute criteria for model-data fit.

• “Let your conscience be your guide”
  – Conduct a variety of analyses, consider the full set of results, and be mindful of these with regard to the intended application of the IRT model.
Next...

- Test Reliability and Development with IRT
- Test Score Equating with IRT