Modeling hint use and response accuracy in learning environments

Maria Bolsinova¹
in collaboration with Benjamin Deonovic², Meirav Attali²,³, Burr Settles⁴, Masato Hagiwara⁴, Alina von Davier² & Gunter Maris¹

¹ - ACTNext, Amsterdam, Netherlands, ² - ACTNext, Iowa City, USA
³ - Fordham University, NYC, USA, ⁴ - Duolingo, Pittsburgh, USA

Maryland Assessment Conference, 2018
Outline

Hints, scaffolds, and adaptive learning systems

Duolingo data

Building the models

Results

Discussion
Zone of proximate development (L. Vygotsky)
Adaptive learning systems

- Adaptive learning systems are designed to dynamically adjust the level or type of learning material based on an individual learner’s abilities or skill attainment (and other characteristics)

- Some features
  - Learner-controlled navigation
  - Interactivity
  - Gamification
  - Transparency

- Monitoring of the development of learners’ skills is crucial to adapt the learning material to their level
Scaffolds in adaptive learning systems

Two ways of giving hints

1. A scaffold/hint/help message is presented if a learner provides an incorrect response
2. A hint can be requested by the learner before providing a response
Hints after an incorrect response

- Whether a hint is provided ($Y_i=1$ if yes, and $Y_i = 0$) is fully determined by the response accuracy on the first attempt to solve the item ($Y_i = 1$ if and only if $X_{i1} = 0$)
- Hints do not provide additional information about ability over and above accuracy on the first attempt
- If a hint was presented, then extra accuracy data would be available (second attempt)
- Polytomous IRT models (i.e., 3 - correct without any hints, 2 - correct after one hint, 1 - correct after two hints, 0 - incorrect), see e.g., Lee, Palazzo, Warnakulasooriya, Pritchard (2008)
Hints on demand

- Learners themselves decide whether to use hints on an item which gives them for freedom and control over their learning process
- Hint use is not directly linked to response accuracy
- Hint use itself might be informative about ability
- Other individual differences between the learners might be also affecting hint use
Duolingo: Adaptive language learning system

- Launched in 2012 (Carnegie Mellon University spinoff)
- More than 200 millions learners globally
- 73 language courses
- Free content
Hints in Duolingo

Write this in English

La biblioteca tiene muchos libros.

Type in English

Skip

USE WORD BANK

Check
Hints in Duolingo

Write this in English

La biblioteca tiene muchos libros.

Type in English

books
Data from newly registered active users between November 9th, 2015 and December 8th, 2015

For each course data from a single platform

Translation items from a foreign language

Only full sentences with at least 3 non-article words

Some items were removed to avoid large overlap between words in the sentences

Items and persons were removed if there were no sufficient observations

Extremely easy items, items with low discrimination, and items with extremely low hint use were removed
Duolingo data sets

- Data set 1: Learning Spanish from English
  - 951 learners
  - 66 items

- Data set 2: Learning English from Portuguese
  - 3250 learners
  - 58 items
Example items

Translate from Spanish to English:

- Yo como arroz con pollo.
- ¿Quién soy yo?
- Él no es vegetariano.
- El verano es una estación.
- Él es un hombre como tú.

Translate from English to Portuguese

- The cat is his.
- I have a tomato and an apple.
- We have a mouse.
- The girl has a mouse.
- Today it is hot.
Hints as process data

Borrowing ideas from response time modeling:

- Process data can be included in the scoring rule for ability such that ability would be estimated based both on product data (accuracy) and process data
- Signed residual time model (Maris & van der Maas, 2012):

\[
S = \sum_{i}(2x_i - 1)(d - t_i),
\]

- Fast-correct is better than slow-correct, but fast-incorrect is worse than slow-incorrect
- An IRT model in which the score is the sufficient statistic for ability can be derived
Item scores based on accuracy and hint use

- $X_i$ - accuracy on the item (1 - correct, 0 - incorrect), $Y_i$ - hint use (1 - at least one hint was used, 0 - no hints were used)

- Four outcomes on each item based on accuracy and hint use, each matching a score

$$S_{pi} = \begin{cases} 
0 & \text{if } X_{pi} = 0, Y_{pi} = 0; \\
1 & \text{if } X_{pi} = 0, Y_{pi} = 1; \\
2 & \text{if } X_{pi} = 1, Y_{pi} = 1; \\
3 & \text{if } X_{pi} = 1, Y_{pi} = 0. 
\end{cases}$$
IRT model derived from the sufficiency of the total score

\[ \sum_i S_{pi} \text{ as a sufficient statistic for the person parameter;} \]

\[ \sum_p S_{pi} \text{ as a sufficient statistic for the item parameter.} \]

\[ \Pr(S_i = s | \theta) = \frac{\exp(s(\theta - \delta_i))}{\sum_{t=0}^{3} \exp(t(\theta - \delta_i))}, \]

\( \theta \) - ability latent variable, \( \delta_i \) - difficulty of item \( i \)
Differences in discriminatory power

- Items might differ in the strength of the relationship between the item score and ability
- Extend the model in the same way as the Rasch model or the Signed-residual-time model have been extended

\[
Pr(S_i = s \mid \theta) = \frac{\exp(s \alpha_i(\theta - \delta_i))}{\sum_{t=0}^{3} \exp(t \alpha_i(\theta - \delta_i))},
\]

\(\alpha_i\) - discrimination parameter of item \(i\)
Conditional accuracy: Model property

\[
\begin{align*}
\Pr(X_i = 1 \mid \theta, Y_i = 1) &= \frac{\exp(\alpha_i(\theta - \delta_i))}{1 + \exp(\alpha_i(\theta - \delta_i))} \\
\Pr(X_i = 1 \mid \theta, Y_i = 0) &= \frac{\exp(3\alpha_i(\theta - \delta_i))}{1 + \exp(3\alpha_i(\theta - \delta_i))}
\end{align*}
\]

- The conditional accuracy functions differ only in discrimination (higher if hints were not used), but not in difficulty.
- The difficulty of the item is the point on the ability scale where all four outcomes are equally likely.
Conditional accuracy: What is found in the data

2PL model is fitted to the accuracy data without hints and to the accuracy data with hints separately

![Graph showing the relationship between hints used and difficulty](image)
Relaxing the model: additional item parameters

Each item has three threshold parameters matching the four outcomes

\[
Pr(S_i = s | \theta) = \frac{\exp(s \alpha_i \theta + \delta_{is})}{\sum_{t=0}^{3} \exp(t \alpha_i \theta + \delta_{it})},
\]

\[
\delta_{is} \equiv 0
\]
Hint use on different items: Model property

- The model predicts a general positive correlation between response accuracies on different items (positive manifold), but does not predict it for hint use on different items.

- $Y_i$ and $Y_j$ on two different items are positively correlated when $X_i = X_j$, but negatively correlated when $X_i \neq X_j$.

$$
Pr(Y_i = 1 \mid X_i = 1, \theta) = \frac{\exp(-\alpha_i \theta + \delta_{i2} - \delta_{i3})}{1 + \exp(-\alpha_i \theta + \delta_{i2} - \delta_{i3})},
$$

is negatively related to ability, while

$$
Pr(Y_j = 1 \mid X_j = 0, \theta) = \frac{\exp(\alpha_j \theta + \delta_{j1})}{1 + \exp(\alpha_j \theta + \delta_{j1})}
$$

is positively related to ability.
Hint use on different items: What is found in the data

Tetrachoric correlations between hint use variables on different items

Histogram

Correlation between \( Y_i \) and \( Y_j \)

Frequency

Correlation between \( Y_i \) and \( Y_j \)
Extending the model: additional source of individual differences

Multidimensional nominal response model (Takane & De Leeuw, 1987; Thissen & Cai, 2016)

\[
\Pr(S_i = s | \theta, \eta) = \frac{\exp(s\alpha_i\theta + \mathcal{I}(s \in \{1, 2\})\lambda_i\eta + \delta_{is})}{\sum_{t=0}^{3} \exp(t\alpha_i\theta + \mathcal{I}(t \in \{1, 2\})\lambda_i\eta + \delta_{it})}
\]

\(\eta\) - extra latent variable accounting for the differences in hint use, \(\lambda_i > 0\) is the loading for this latent variable

<table>
<thead>
<tr>
<th>Scores for (\theta)</th>
<th>Scores for (\eta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect w/o hints</td>
<td>0</td>
</tr>
<tr>
<td>Incorrect w hints</td>
<td>1</td>
</tr>
<tr>
<td>Correct w hints</td>
<td>2</td>
</tr>
<tr>
<td>Correct w/o hints</td>
<td>3</td>
</tr>
</tbody>
</table>

Bolsinova et al. (ACTNext)
Alternative scoring rules

1. Incorrect with hints is better than incorrect without a hint
2. Incorrect responses with and without a hint are the same
3. Incorrect without hints is better than asking for a hint
### Alternative scoring rules

1. Incorrect with hints is better than incorrect without a hint
2. Incorrect responses with and without a hint are the same
3. Incorrect without hints is better than asking for a hint

### Scores for ability dimensions

<table>
<thead>
<tr>
<th></th>
<th>Alternative 1</th>
<th>Alternative 2</th>
<th>Alternative 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect w/o hints</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Incorrect w hints</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Correct w hints</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Correct w/o hints</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Alternative approach: IRTrees

Node 1: Hints used?

- Yes:
  - Node 2: Response correct?
    - Yes: $Y_i = 1, X_i = 1$
    - No: $Y_i = 1, X_i = 0$

- No:
  - Node 3: Response correct?
    - Yes: $Y_i = 0, X_i = 1$
    - No: $Y_i = 0, X_i = 0$
IRTree for hint use

- Probabilities at each node are modeled with the 2PL

\[
\begin{align*}
\text{Pr}(X_i = 0, Y_i = 0 \mid \theta, \eta) &= \frac{1}{1 + \exp(\lambda_i \eta + \gamma_i)} \quad \frac{1}{1 + \exp(\alpha_0 \theta_0 + \beta_0)} \\
\text{Pr}(X_i = 0, Y_i = 1 \mid \theta, \eta) &= \frac{\exp(\lambda_i \eta + \gamma_i)}{1 + \exp(\lambda_i \eta + \gamma_i)} \quad \frac{1}{1 + \exp(\alpha_1 \theta_1 + \beta_1)} \\
\text{Pr}(X_i = 1, Y_i = 1 \mid \theta, \eta) &= \frac{\exp(\lambda_i \eta + \gamma_i)}{1 + \exp(\lambda_i \eta + \gamma_i)} \quad \frac{\exp(\alpha_1 \theta_1 + \beta_1)}{1 + \exp(\alpha_1 \theta_1 + \beta_1)} \\
\text{Pr}(X_i = 1, Y_i = 0 \mid \theta, \eta) &= \frac{1}{1 + \exp(\lambda_i \eta + \gamma_i)} \quad \frac{\exp(\alpha_0 \theta_0 + \beta_0)}{1 + \exp(\alpha_0 \theta_0 + \beta_0)}
\end{align*}
\]

- Potentially different latent variables might be active depending on the outcome of Node 1, and also the item parameters might also be different, otherwise constraints may be imposed \( \theta_0 = \theta_1, \alpha_0 = \alpha_1, \beta_0 = \beta_1 \)
For models with single $\delta_i$ we wrote EM-algorithm in R, all other models we estimated using R-package mirt

- **Divide-by-total models**
  1. {0123} scores for $\theta$, single $\delta$
  2. {0123} for $\theta$, 3 $\delta_i$s
  3. {0123} for $\theta$, [0110] for $\eta$, 3 $\delta_i$s
  4. {1023} for $\theta$, [0110] for $\eta$, 3 $\delta_i$s
  5. {0012} for $\theta$, [0110] for $\eta$, 3 $\delta_i$s
  6. {2013} for $\theta$, [0110] for $\eta$, 3 $\delta_i$s

- **IRTree models**
  1. $\theta_0 = \theta_1$, $\alpha_{0i} = \alpha_{1i}$, $\beta_{0i} = \beta_{1i}$
  2. $\theta_0 = \theta_1$, $\alpha_{0i} \neq \alpha_{1i}$, $\beta_{0i} = \beta_{1i}$
  3. $\theta_0 = \theta_1$, $\alpha_{0i} = \alpha_{1i}$, $\beta_{0i} \neq \beta_{1i}$
  4. $\theta_0 = \theta_1$, $\alpha_{0i} \neq \alpha_{1i}$, $\beta_{0i} \neq \beta_{1i}$
  5. $\theta_0 \neq \theta_1$, $\alpha_{0i} \neq \alpha_{1i}$, $\beta_{0i} \neq \beta_{1i}$
## 10-fold cross validation

<table>
<thead>
<tr>
<th>Model</th>
<th>LL in testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scoring-rule-based models</strong></td>
<td></td>
</tr>
<tr>
<td>$a = [0, 1, 2, 3]$, no $\alpha_i$, single $\delta_i$</td>
<td>-14627.31</td>
</tr>
<tr>
<td>$a = [0, 1, 2, 3]$, single $\delta_i$</td>
<td>-14557.70</td>
</tr>
<tr>
<td>$a = [0, 1, 2, 3]$, three $\delta_i$s</td>
<td>-13141.73</td>
</tr>
<tr>
<td>$a = [0, 1, 2, 3]$, $b = [0, 1, 1, 0]$, three $\delta_i$s</td>
<td>-11924.42</td>
</tr>
<tr>
<td>$a = [1, 0, 2, 3]$, $b = [0, 1, 1, 0]$, three $\delta_i$s</td>
<td>-11933.00</td>
</tr>
<tr>
<td>$a = [0, 0, 1, 2]$, $b = [0, 1, 1, 0]$, three $\delta_i$s</td>
<td>-11921.60</td>
</tr>
<tr>
<td>$a = [2, 0, 1, 3]$, $b = [0, 1, 1, 0]$, three $\delta_i$s</td>
<td>-11908.37</td>
</tr>
<tr>
<td><strong>IRTree models</strong></td>
<td></td>
</tr>
<tr>
<td>$\theta_0 = \theta_1$, $\alpha_{0i} = \alpha_{1i}$, $\beta_{0i} = \beta_{1i}$</td>
<td>-12001.43</td>
</tr>
<tr>
<td>$\theta_0 = \theta_1$, $\alpha_{0i} \neq \alpha_{1i}$, $\beta_{0i} = \beta_{1i}$</td>
<td>-11951.00</td>
</tr>
<tr>
<td>$\theta_0 = \theta_1$, $\alpha_{0i} = \alpha_{1i}$, $\beta_{0i} \neq \beta_{1i}$</td>
<td>-11970.82</td>
</tr>
<tr>
<td>$\theta_0 = \theta_1$, $\alpha_{0i} \neq \alpha_{1i}$, $\beta_{0i} \neq \beta_{1i}$</td>
<td>-11943.81</td>
</tr>
<tr>
<td>$\theta_0 \neq \theta_1$, $\alpha_{0i} \neq \alpha_{1i}$, $\beta_{0i} \neq \beta_{1i}$</td>
<td>-11941.90</td>
</tr>
</tbody>
</table>
Some additional results

- The item slopes on the additional dimension were rather strong - mean of 1.69, ranging from 0.59 to 2.76
- $\theta$ and $\eta$ were not correlated (estimate of .05)
Discussion

- Different strategies for joint modeling of hint use and accuracy
- Hint use is informative of ability
- Hint use depends not only on ability but also on some additional personal characteristics
- Further research into the person predictors for the hint-use latent variable is needed
- Information about learners’ tendency to use or not use hints may be used in the adaptive learning systems to give additional feedback to students and customize their learning paths
- Response times or other process data may also be included in the model using similar modeling strategies
Thank you!

For additional information: maria.bolsinova@act.org