Estimating learners' ability in a learning platform with options for learners' choices

Meirav Arieli-Attali

January 2021

הכינוס ה-17 של האגודה הישראלית לפסיכומטריקה 2021

*Study done while working at ACTNext

Adaptive Learning and Assessment Platforms

- In order to be adaptive – we need to assess learners' performance
- Digital learning platform collect performance data by default
- Can we assess learners' ability and knowledge
  - Without having to pause learning to take a test?
- How much is the data collected by default indicative of ability?
  - How does the behavioral data (e.g., leaners choices) interfere or affect ability estimates? ➔ How messy is the data

The challenge -

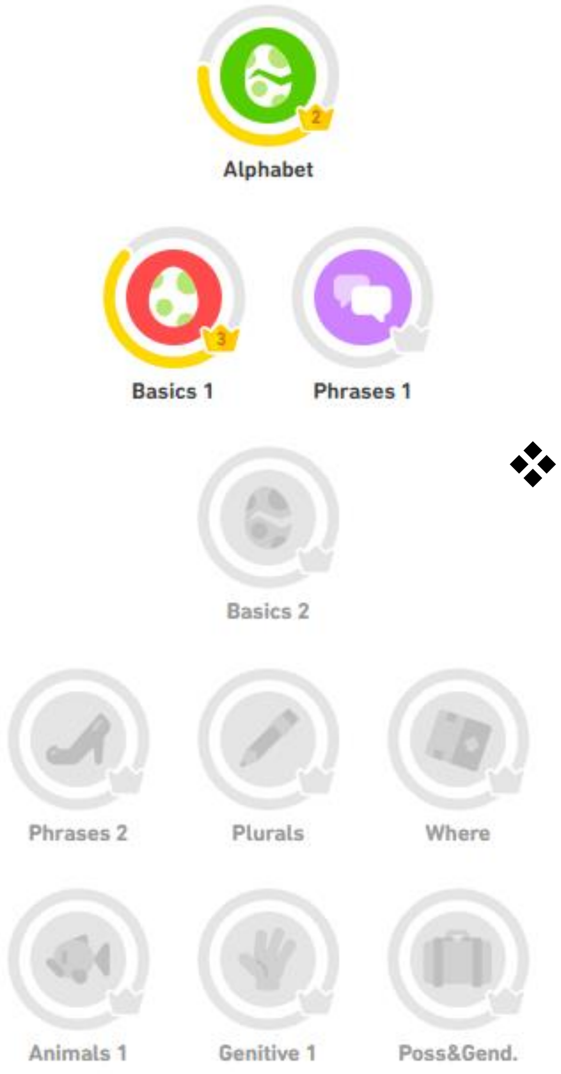
- Learning platform include features not suitable to be used for measurement
  - Missing data
  - Multiple attempts
  - Hints
  - Feedback
  - Learners’ choices (may depend on system reward system)
  - Not standardized use (learners may take breaks, be interrupted…)

The data -

- Duolingo, an online language learning platform with more than 200 million registered users
- organized into lessons, each is a set of questions (=items) with immediate correctness feedback

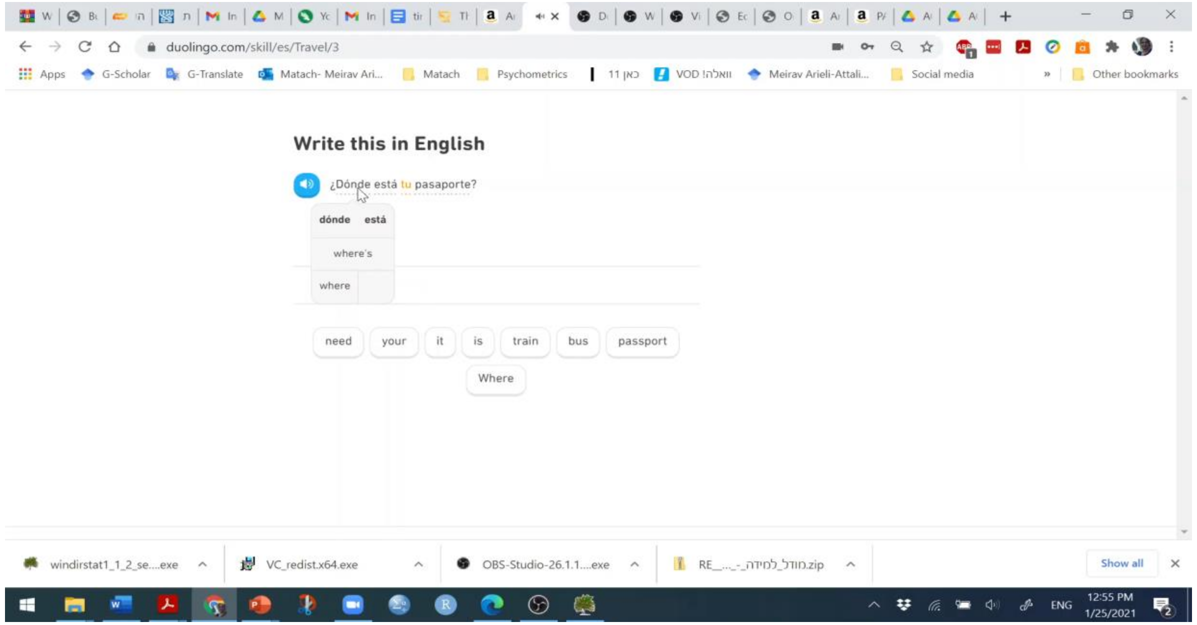
- Item type with a HINT option: translate a sentence from the learning language into known language
  - hovering-over a word in the sentence opens-up a pop-up with the translation of that word
  - Learners could hover-over **one** or **all words** to see their translation
  - This is a subtle “hint request”

The Duolingo App organized into lessons….



❖ Lessons / rows are unlocked with progress

The hover-over serves as a hint



Data Building & Cleaning

- Two data sets:
  - Spanish-from-English
  - English-from-Portuguese
- By date: items completed between November 9, 2015, and December 8, 2015.
- Learners: only new accounts that reached at least the tenth row (>10 lessons)
- Platform: only data from a single platform
  - Android in the Spanish-from-English
  - iOS was used for the English-from-Portuguese

Items:

- Only the first time a learner responded to an item (repeated items were excluded)
- Only items with complete sentences (i.e., not word combinations or single words)
- With less than 70% overlap of words with other items (to enable the independence assumption)

Resulted in

Spanish-from-English – 89 items and 1109 learners
English-from-Portuguese - 99 items and 3845 learners

Methodology -

- examined several models jointly modeling response accuracy and hint use
  - Inspired by the signed-residual-time model (Maris & van der Maas, 2012)
- used two datasets, one for developing the models, the second to apply and choose the best fitting model
- used extension of IRT-family models

Assumption:

information on whether learners use hints or not can be used to obtain additional information about the measured abilities or skills -➔ there is construct relevant information in the choice to use a hint

The scoring models

based on both

- whether the response was correct - *Xpi*
- whether it was obtained with a hint - *Ypi*

$$S_{pi} = \begin{cases} 0 \text{ if } X_{pi} = 0, Y_{pi} = 0; \\ 1 \text{ if } X_{pi} = 0, Y_{pi} = 1; \\ 2 \text{ if } X_{pi} = 1, Y_{pi} = 1; \\ 3 \text{ if } X_{pi} = 1, Y_{pi} = 0. \end{cases}$$

Similarly to the signed-residual-time model, correct responses without hints are encouraged, while incorrect response without hints are discouraged by the scoring rule

Ability Estimate Models

- IRT models can be derived from this scoring rule

Rasch / 1PL model

$$\Pr(S_i = s \mid \theta) = \frac{\exp(s(\theta - \delta_i))}{\sum_{r=0}^{3} \exp(r(\theta - \delta_i))},$$

where $s \in \{0, 1, 2, 3\}$, θ is ability latent variable, and δ_i is the difficulty of item i.

Note: This is a constrained version of the partial credit model in which there is a single item difficulty parameter instead of multiple threshold parameters.

2PL model

$$\Pr(S_i = s \mid \theta) = \frac{\exp(s\alpha_i(\theta - \delta_i))}{\sum_{r=0}^{3} \exp(r\alpha_i(\theta - \delta_i))}$$

where $\alpha_i > 0$ is the discrimination parameter of item i.

2PL model + several difficulty parameters

$$\Pr(S_{pi} = s \mid \theta) = \frac{\exp(s\alpha_i\theta + \delta_{is})}{\sum_{r=0}^{3} \exp(r\alpha_i\theta + \delta_{ir})},$$

where δ_{is} is a category-specific parameter with δ_{i0} being constrained to be equal to zero

This is actually
The generalized partial credit model
Muraki (1992)

BUT.....

- We noticed that hint use variables were correlated....

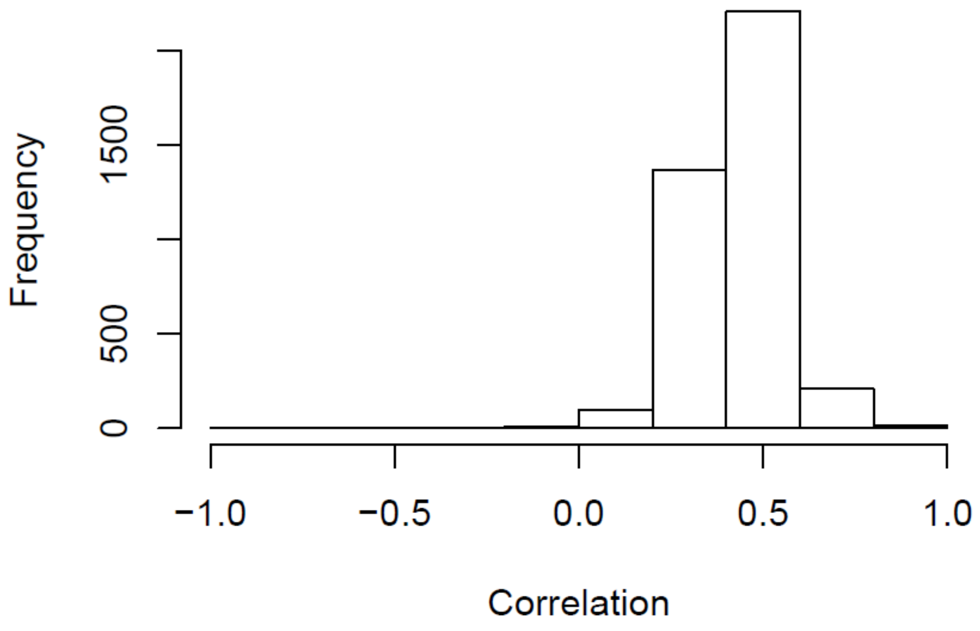


Figure 2. Histogram of the tetrachoric correlations between hint use variable on different items in the Spanish-from-English Duolingo data set.

So we added a new variable ....

- New variable: *Tendency-To-Use-Hint - η*

$$\Pr(S_i = s \mid \theta, \eta) = \frac{\exp(s\alpha_i\theta + I(s \in \{1,2\})\lambda_i\eta + \delta_{is})}{\sum_{r=0}^{3}\exp(r\alpha_i\theta + I(r \in \{1,2\})\lambda_i\eta + \delta_{ir})}$$

where η is the extra latent variable accounting for the differences in hint use, $\lambda_i > 0$ is the item loading for this latent variable, and I(condtion) is the identity function which takes a value of one if the condition is satisfied, and a value of zero if it is not.

This is actually
The multidimensional nominal response model
(Takane & De Leeuw, 1987; Thissen & Cai, 2016)

Additional models

$$S_{pi} = \begin{cases} 0 \text{ if } X_{pi} = 0, Y_{pi} = 0; \\ 1 \text{ if } X_{pi} = 0, Y_{pi} = 1; \\ 2 \text{ if } X_{pi} = 1, Y_{pi} = 1; \\ 3 \text{ if } X_{pi} = 1, Y_{pi} = 0. \end{cases}$$

Original is **[0, 1, 2, 3]** ➔ meaning: use of hint is a resource

Other options:

For incorrect responses

- without hints can be considered better than with hints **[1, 0, 2, 3]**
- no difference with and without hints **[0, 0, 1, 2]**

Hint use reflect lower ability (confidence in ability?)

- Incorrect responses without hints are better than the responses with hints regardless of correctness. **[2, 0, 1, 3]**

Ignore hint use / traditional scoring

- Only correct responses without hints receive full credit, while all other options receive no credit **[0, 0, 0, 1]**

Results

Model	npar	AIC	BIC	CVLL
Scoring-rule-based models				
$IH_- < IH_+ < CH_+ < CH_-$, no α_i, single δ_i, no η	100	275065	275621	-137432
$IH_- < IH_+ < CH_+ < CH_-$, single δ_i, no η	198	273548	274649	-136867
$IH_- < IH_+ < CH_+ < CH_-$, no η	396	241118	243320	-120584
$IH_- < IH_+ < CH_+ < CH_-$	496	210563	213322	-105273
$IH_+ < IH_- < CH_+ < CH_-$	496	210622	213381	-105304
$(IH_-, IH_+) < CH_+ < CH_-$	496	210522	213280	-105254
$IH_+ < CH_+ < IH_- < CH_-$	496	210653	213412	-105327
$(IH_-, IH_+, CH_+) < CH_-$	496	210754	213512	-105361

With η

Ability & Tendency to Use Hints

In the selected scoring-rule-based **[0, 0, 1, 2]**

- correlation equal to **.13** [*CI* : .09, .17].
  - more able students are slightly more likely to use hints
- individual differences in the tendency to use hints was larger than individual differences in ability

➔What does this variable of "tendency-to-use-hints" actually mean?

- Use hint as a learning tool
- Learners don't want to err (error is penalized)

Summary & Discussion

- We showed a way to analyze data, taking into account variability in learners' behavior – here: HINT USE
  - We needed to do a lot of cleaning to the data ahead of all analyses
  - We added a behavioral factor – the tendency to use hint
- Hint use may be perceived conceptually as "partial knowledge"
- Question of validity -
  - What is the validity of these ability scores? What do we gain from estimating ability in this way?
- Would we get the same preferred model if learners knew their ability is estimated while working in the system?
- How do our results depend on the specific system and its reward system?

Research Team (ACTNext/ Duolingo)

Meirav Arieli-Attali

Maria Bolsinova

Benjamin Deonovic

Burr Settles

Masato Hagiwara

Gunter Maris

Alina von Davier