Evaluation of Educational Systems

Teaching for Learning Core Lecture 7

Matthieu Brinkhuis

October 18, 2018

Utrecht University, Information and Computing Sciences
m.j.s.brinkhuis@uu.nl
Today’s lecture:

- Introduction on evaluation
- Evaluating a Large-Scale Online Learning System (Brinkhuis et al., 2018)
  1. Layered evaluation of adaptive educational systems (Brusilovsky, Karagiannidis, and Sampson, 2004)
  2. Student model quality evaluation (Pelánek, 2015)
  3. Evaluating large scale surveys (Zwitser, Glaser, and Maris, 2016)
  4. Evaluating adaptive educational systems with learning curves (Martin et al., 2011)
  5. Evaluation of validity (Wools, 2010)
By the end of this lecture, you’ll be able to:

- Understand different types of evaluation for educational systems
- Apply selected evaluation principles
- Analyze an existing evaluation by discussing parts related to this lecture and the literature
1. Introduction

2. Evaluating a Large-Scale Online Learning System

3. Papers for presentations

4. Closing
Introduction
Generally in measurement, two concepts are deemed important in evaluation:

- Reliability
- Validity

Important to distinguish how these are important in the context of modeling (prediction or inference: INFOMDSS, see also Lazer et al. (2014) for a discussion on how these are crucial in the age of big data).
Statistical learning

\[ Y = f(X) + \epsilon \]  

where \( f \) is unknown function of \( X_1, X_2, \ldots, X_p \) and \( \epsilon \) is a random error term, independent of \( X \) (James et al., 2013).

Estimate \( \hat{f} \) for:

- prediction (black box)
- inference (interest in associations, what is the type of relationship, etc.)

There is a trade-off between prediction accuracy and model interpretability (Waa et al., 2018; James et al., 2013).
Evaluating a Large-Scale Online Learning System
“[...] in the near future it will be possible to continuously assess and store the unfolding life history (trajectory in behavior space) of each individual” (Molenaar, 2004).

It should be possible to create completely individualized educational programs, by teachers that are fully informed on the progress and learning difficulties of each student.

Contribute to this debate by providing a case study (Klinkenberg, Straatemeier, and Maas, 2011).
Figure 1: Mathgarden landing page: the garden.
Figure 2: Mathgarden example item, here with multiple choice responses.
The Math Garden in numbers (early 2018):

- over 853,300,000 responses
- over 452,000 K-12 children
- 5,300 schools and many more household subscriptions
- 26 arithmetic domains
- totaling more than 37,000 different items
- rate about 900,000 per school day

Not suitable for classical psychometrics or statistics.
Figure 3: Cohort development over 3 years.
Scoring rule

\[ S_{pi} = (2x_{pi} - 1)(d - t_{pi}), \tag{2} \]

where \( S_{pi} \) denotes the score earned by user \( p \), \( d \) denotes the time limit and \( x_{pi} \in \{0, 1\} \) and \( t_{pi} \in [0, d] \) denote the accuracy and the response time of user \( p \) on item \( i \). (Maris and Maas, 2012)

**Figure 4:** Response time model.
Adaptive Item Selection

\[ \theta_p \rightarrow \theta_p + K(S_{pi} - E(S_{pi})) , \]  
\[ \delta_i \rightarrow \delta_i - K(S_{pi} - E(S_{pi})) , \]  

where \( K \) is a scaling factor and the expected score \( E(S_{pi}) \) is based on the current ability estimate \( \theta_p \) and item difficulty estimate \( \delta_i \):

\[ E(S_{pi}|\theta_p, \delta_i) = d \frac{\exp (2d(\theta_p - \delta_i)) + 1}{\exp (2d(\theta_p - \delta_i)) - 1} - \frac{1}{\theta_p - \delta_i} \]  

where \( d \) is the time limit (Klinkenberg, Straatemeier, and Maas, 2011).
We will evaluate several measures from the methodology underlying the Math Garden:

• Evaluate prediction accuracy
• Evaluate reliability of (item) difficulty estimates
• Diagnose model misfit in local strategies
• Diagnose model misfit in scoring rule
Figure 5: Model fit (O-E) for a single person over time.
Figure 6: Model fit (O-E) for all responses.
Figure 7: Reliability of mirror item ratings.
Figure 8: Example balance scale item.
**Diagnosis of Model Misfit**

**Figure 9:** Local item strategies interacting with the adaptive model.
Figure 10: Examples of development over time, not all as expected.
Many opportunities in today’s educational data, but how to move forward?
Two main considerations for the future of learning analytics.
Learning analytics must actively help direct a student towards his or her educational objective (e.g., such as effortful practice on the level of the individual child)

- embedded learning analytics: dynamically steer learning experience in the desired direction.
- experimentation: supervision (A/B)
- theory building: pedagogical interventions
Large-scale online learning systems are increasingly:

- invasive
- complex
- dynamic
- goal-directed

Dynamics not completely understood: research responsibility.
Learning inherently difficult to track, let alone to deliberately navigate towards a desired goal: active analytics needed.

Educational big data provides a window into the complexity and dynamics of learning in vivo: opportunities for research.
Papers for presentations
Layered evaluation of adaptive educational systems

Why evaluate adaptive learning systems as a whole?

The paper by Brusilovsky, Karagiannidis, and Sampson (2004) is focused on *layered* evaluation, such as user models and adaptive decision making.

<table>
<thead>
<tr>
<th>Adaptive system</th>
<th>Traditional evaluation approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive hypermedia</td>
<td>Measure of recall and precision, similarity/relevance metrics, comparison of the system with and without adaptation</td>
</tr>
<tr>
<td>Plan recognition</td>
<td>Percentage of actual plans recognized in a test corpus of plans; frequency and accuracy of predicted actions, comparison with an expert human recognizer</td>
</tr>
<tr>
<td>Mixed-initiative interaction</td>
<td>Comparing system responses choices with human choices, efficiency of the dialogue needed to achieve an information transfer task with either human-human dialogues or with theoretically minimum dialogues</td>
</tr>
<tr>
<td>User interfaces</td>
<td>Subjective user satisfaction, task completion speed, error rate - quality of task achievement</td>
</tr>
</tbody>
</table>

*Table 1:* Traditional ways of evaluation of adaptive systems.
Evaluating quality of student models

The paper by Pelánek (2015) discusses several metrics for evaluating student models, and relates them. He makes a difference between probabilistic and qualitative metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>$\frac{1}{n}\sum_{i=1}^{n}</td>
</tr>
<tr>
<td>RMSE</td>
<td>$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2}$</td>
</tr>
<tr>
<td>LL</td>
<td>$\sum_{i=1}^{n} o_i \log(p_i) + (1 - o_i) \log(1 - p_i)$</td>
</tr>
</tbody>
</table>

Table 2: Most often used metrics based on probabilistic understanding of errors (p.4).
How can we in a meaningful way compare countries in international surveys?

The paper by Zwitser, Glaser, and Maris (2016) is focused on DIF in surveys, and evaluates surveys by a new look at DIF.

Since the paper is rather technical, tip for the presenter: focus on the evaluation aspect, not the method.

Figure 11: 95% rank intervals per country.
How can we evaluate the validity of a learning system?

The paper by Wools (2010) is focused on validating by the argument-based approach. It forces you to provide argumentations on your claims, and is closely connected to the goal and use of your system.

Figure 12: Toulmin’s model for arguments.
Closing
By the end of this lecture, you’ll be able to:

- Understand different types of evaluation for educational systems
- Apply selected evaluation principles
- Analyze an existing evaluation by discussing parts related to this lecture and the literature
Thank you.


