What Does It Take to Optimize Human Learning?

Jacob Whitehill
Worcester Polytechnic Institute
jrwhitehill@wpi.edu

TU-Delft Data Science Seminar (31 March 2017)
My research

• How machine learning and computer vision can predict and optimize human learning outcomes:
  • Automatic estimation of students’ facial expressions
  • Reinforcement learning approaches to optimized teaching
  • Educational data mining from MOOC and ITS datasets
Theme of this talk

• This talk will draw inspiration from how “teaching” can be framed as an **optimal control** problem (e.g., Suppes 1964; Matheson 1964):

  • How should the teacher act at each moment in time, based on feedback from the student, so as to maximize some objective function?

  • In the language of optimal control, we can define “teaching” in terms of **actions**, **states**, and **feedback** (observations).
Optimized teaching

- Teacher’s action $A_t \in A$ can impact student’s state:
Optimized teaching

• Teacher’s **action** $A_t \in A$ can impact student’s **state**: 

 Assigning homework on topic X can help student increase student’s skill on X.

Teacher

Student

Actions $A_1, \ldots, A_t$
Optimized teaching

- Teacher’s action $A_t \in A$ can impact student’s state:
  - Giving problems that are too difficult can cause student to become frustrated.

---

Teacher

Student

Actions $A_1, \ldots, A_t$

Knowledge state
Affective state
Optimized teaching

- Student can give **feedback** to the teacher:
Optimized teaching

- Student can give **feedback** to the teacher:
  - Answers to practice problems.

![Diagram](image.png)

- Knowledge state
- Affective state

**Feedback**

**Actions**

$O_1, \ldots, O_t$

$A_1, \ldots, A_t$
Optimized teaching

- Student can give **feedback** to the teacher:
  - Facial expression indicative of emotional state.

\[
O_1, \ldots, O_t \quad \text{Feedback}
\]
\[
A_1, \ldots, A_t \quad \text{Actions}
\]

Teacher

Student

Knowledge state
Affective state
Optimized teaching

- Implicitly, teachers must solve a **control problem**: what action $A_{t+1}$ to take next, given history of feedback & prior actions?

**Decision function:**

$$A_{t+1} = f(A_1, \ldots, A_t, O_1, \ldots, O_t)$$

**Feedback**

$O_1, \ldots, O_t$

**Teacher**

Should meet those who

**Student**

Knowledge state
Affective state

Prior actions

Feedback
Optimized teaching

• One approach to tackling this control problem to use **model-based control** to decompose the control problem into several components…
Model-based control for optimized teaching

1. Create a model of how the student changes state given the teacher’s action (i.e., \textbf{dynamics model}): 
\[ P(s_{t+1} \mid s_t, a_t) \]

- E.g., dynamics model might express that MOOC learners in “casually interested” state may transition to “dropped out” state if the MOOC assigns too much work.
Model-based control for optimized teaching

2. Create a model of how student expresses her/his state (i.e., observation model):

\[ P(o_t \mid s_t, a_t) \]

• E.g., learners who have mastered a skill tend to do very well on the test (obvious).
Model-based control for optimized teaching

2. Create a model of how student expresses her/his state (i.e., observation model):

\[ P(o_t \mid s_t, a_t) \]

- E.g., learners who are in a “frustrated” state may click the same button many times in a row.
Model-based control for optimized teaching

3. Decide which states are more preferable via function $r$, e.g.:

- We prefer the student to know a skill than not to know the skill.
- We prefer the student to be “very engaged” over “very frustrated”.
Model-based control for optimized teaching

• Given:
  • Dynamics model
  • Observation model
  • Preference function $r$

• We can employ machine learning algorithms to find the decision function $f$ that (approximately) optimizes the expected long-term preference of student states:

$$ f = \arg \max_{f'} \mathbb{E} \left[ \sum_{t=1}^{T} r(S_t) \mid f' \right] $$
Model-based control for optimized teaching

- **Analogy**: To make an airplane fly, having a model of how the plane responds to the pilot’s actions is very useful.

- Don’t need to **randomly** “twiddle” with the yoke & throttles until we stumble onto the right inputs.

- Don’t need to **manually** decide how to respond to every possible flight scenario.

- Given the dynamics & observation models, we can use machine learning & reinforcement learning to help us to compute the decision function.
Overview

• I will present 3 projects:
  1. Automatic Measurement of Student Engagement From Video
  2. Predicting Dropout among MOOC Learners
  3. Crowdsourcing 399 Tutorial Videos on Logarithms
Overview

• I will present 3 projects:

1. Automatic Measurement of Student Engagement From Video
   Observation model of engaged students: \( P(o_t \mid s_t, a_t) \)

2. Predicting Dropout among MOOC Learners
   Dynamics model of if/when MOOC learners drop out: \( P(s_{t+1} \mid s_t, a_t) \)

3. Crowdsourcing 399 Tutorial Videos on Logarithms
   Grow the action space from \( \mathcal{A} \) to \( \mathcal{A}' \supset \mathcal{A} \)
Automatic Measurement of Student Engagement from Video

\[ P(o_t \mid s_t, a_t) \]
Research questions

1. How accurately can we estimate student engagement automatically using computer vision?

2. What information relevant to teaching and learning is expressed in the face?
Student engagement

• Starting around 1980s, “engagement” has emerged as a key metric of teaching success (e.g., Larson & Richards 1991).

• At a recent Harvard conference on learning & teaching, “engagement” was the word that professors most strongly associated with effective teaching.

Whitehill, Serpell, et al. 2014
Conventional methods of measuring student engagement

- Surveys and observational checklists:
  - Primacy/recency effects
  - Reluctance to give honest responses
  - Lack of temporal resolution

- Physiological sensors (EEG, GSR):
  - Obtrusive; requires physical contact

Whitehill, Serpell, et al. 2014
Measuring student engagement

• Our goal: develop automatic detector of student engagement using web-camera & computer vision.

• Specifically, estimate from video how “engaged” the student appears to an average observer.

• Unobtrusive

• High temporal resolution.
Dataset

• 30 undergraduate students interacting with an iPad-based intelligent tutoring system.
Dataset

• Each student in the dataset followed the protocol:
  • **Pre-test** on a cognitive skills task ("Set")
  • 30 minutes of **training** on the task
  • **Post-test**
Our approach

• First step: define what “engagement” means.

• Based on video data, we defined four engagement levels:
  • 1 = Not engaged at all.
  • 2 = Nominally engaged.
  • 3 = Engaged.
  • 4 = Very engaged.
Examples

Not engaged at all

Labels rated Engagement=1

Whitehill, Serpell, et al. 2014
Examples

Nominally engaged

Labels rated Engagement=2

Whitehill, Serpell, et al. 2014
Examples

Engaged

Labels rated Engagement=3

Whitehill, Serpell, et al. 2014
Examples

Very engaged

Labels rated Engagement=4

Whitehill, Serpell, et al. 2014
Labeling

• We labeled ~20,000 face images from the 30 participants in the dataset.

• Given images+labels, we can use machine learning to create an automatic engagement detector.
Recognition architecture

Input image

48x48 face image

Neural network

Gabor filters

Estimate of “engagement”

Whitehill, Serpell, et al. 2014
System accuracy

• Accuracy in distinguishing E=4 from E=1: 93%
System accuracy

- Accuracy on fine-grained classification tasks (e.g., “Is this student in engagement level 2?”):
  - Automated detector: 73%
  - Human-human: 72%

Whitehill, Serpell, et al. 2014
Recognition example
Recognition example

Whitehill, Serpell, et al. 2014
Recognition example
Correlations with task performance

• We also investigated whether engagement is correlated with students’ test performance.

• “Engagement” was positively correlated with both pre-test and post-test:

  • $\text{corr(Engagement, Pre-test)}: 0.57$
  • $\text{corr(Engagement, Post-test)}: 0.47$
Correlations with task performance

• We also investigated whether engagement is correlated with students’ test performance.

• “Engagement” was positively correlated with both pre-test and post-test:

• \[ \text{corr(Engagement, Pre-test): 0.57} \]
  \[ \text{corr(Engagement, Post-test): 0.47} \]
  \[ \text{corr(Pre-test, Post-test): 0.44} \]
Correlations with task performance

• In other words:

  • To predict a student’s post-test score, the average engagement level is just as informative as the pre-test score.
What else does the face reveal?

- How is the student doing?
- Is the task too easy? Too hard?
- Is the student trying to succeed?
What else does the face reveal?
Predicting Dropout among MOOC Learners

\[ P(s_{t+1} \mid s_t, a_t) \]
Completion and stop-out

- As of March 2017:
  - ~3 million students have registered in 184 HarvardX courses and generated >1 billion clickstream events.
  - >74K students earned certificates.
  - Course completion rates are low: ~4%
  - Not necessarily a problem: since there is 0 cost, students get what they want and leave.
Completion and stop-out

• However:
  
  • Completion rates are low (22%) even for students who express the *intent* to complete in a pre-course survey (Reich 2014).
  
  • Some students might want to learn more but *disengage* due to boredom, frustration, etc.
  
  • Some students could be “nudged” to persevere.

Whitehill, Reich, et al. 2015
Whitehill, et al. 2017
MOOC dropout prediction

• Since ~2012, researchers have begun developing algorithms to predict which MOOC learners will drop out based on clickstream, discussion forum, and social network features:

<table>
<thead>
<tr>
<th>Study</th>
<th>#MOOCs</th>
<th>Features</th>
<th>Architecture</th>
<th>Training setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balakrishnan &amp; Coetzee [1]</td>
<td>1</td>
<td>Clickstream</td>
<td>HMM + SVM</td>
<td>Same course</td>
</tr>
<tr>
<td>Boyer &amp; Veeramachaneni [2]</td>
<td>3</td>
<td>Clickstream</td>
<td>TL+LR</td>
<td>Different offering</td>
</tr>
<tr>
<td>Coleman et al. [5]</td>
<td>1</td>
<td>Clickstream</td>
<td>LDA+LR</td>
<td>Same course</td>
</tr>
<tr>
<td>Crossley et al. [6]</td>
<td>1</td>
<td>Clickstream; NLP</td>
<td>DFA</td>
<td>Same course</td>
</tr>
<tr>
<td>Fei &amp; Yeung [7]</td>
<td>2</td>
<td>Clickstream</td>
<td>RNN</td>
<td>Same course</td>
</tr>
<tr>
<td>He et al. [10]</td>
<td>2</td>
<td>Clickstream</td>
<td>Smoothed LR</td>
<td>Different offering</td>
</tr>
<tr>
<td>Jiang et al. [12]</td>
<td>1</td>
<td>Social network; grades</td>
<td>LR</td>
<td>Same course</td>
</tr>
<tr>
<td>Kizilcec et al. [14, 9]</td>
<td>20</td>
<td>Clickstream</td>
<td>SVM</td>
<td>Different course</td>
</tr>
<tr>
<td>Kloft et al. [15]</td>
<td>1</td>
<td>Clickstream</td>
<td>LR</td>
<td>Same course</td>
</tr>
<tr>
<td>Koedinger et al. [16]</td>
<td>1</td>
<td>Clickstream; grades</td>
<td>LR</td>
<td>Same course</td>
</tr>
<tr>
<td>Robinson et al. [17]</td>
<td>1</td>
<td>Survey; NLP</td>
<td>LR</td>
<td>Same course</td>
</tr>
<tr>
<td>Rose et al. [25, 18]</td>
<td>1</td>
<td>Forum; social network</td>
<td>SA</td>
<td>Same course</td>
</tr>
<tr>
<td>Stein &amp; Allione [20]</td>
<td>1</td>
<td>Clickstream; survey</td>
<td>SA</td>
<td>Same course</td>
</tr>
<tr>
<td>Taylor et al. [21]</td>
<td>1</td>
<td>Clickstream</td>
<td>LR</td>
<td>Same course</td>
</tr>
<tr>
<td>Whitehill et al. [22]</td>
<td>10</td>
<td>Clickstream</td>
<td>LR</td>
<td>Different course</td>
</tr>
<tr>
<td>Xing et al. [24]</td>
<td>1</td>
<td>Clickstream; social network</td>
<td>PCA+{BN,DT}</td>
<td>Same course</td>
</tr>
<tr>
<td>Ye &amp; Biswas [26]</td>
<td>1</td>
<td>Clickstream</td>
<td>LR</td>
<td>Same course</td>
</tr>
</tbody>
</table>
MOOC dropout prediction

• Dropout predictors could facilitate automated MOOC interventions by identifying who needs an intervention and when they need it.

• Common architecture:
  • For each week \( w \) of the course:
    • Compute feature vector \( f_w \) for each student using data up through week \( w \).
    • Use trained model to estimate \( P(\text{DropOut} \mid f_w) \).
MOOC dropout prediction

• To-date, there have been only a few studies (Kizilcec, et al. 2015; Whitehill, Reich, et al. 2015) that explore how to actually use a trained dropout detector to conduct an intervention.

• When deploying a MOOC dropout detector, there is an interesting trade-off between detection accuracy and intervention effectiveness.
Intervention trade-off

• The longer we wait, the more confident we become that the student has truly stopped out.
Classifier output vs. 
# weeks since stop-out

Average Classifier Output versus Days Since Stopout

$P(\text{DropOut} \mid f_w)$

Confidence of dropout detector increases with time since dropout.

Whitehill, Reich, et al. 2015
Whitehill, et al. 2017
Intervention trade-off

• The longer we wait, **the more confident** we become that the student has truly stopped out.

• The longer we wait, **the less likely** the student will respond to the intervention.
Intervention response rates

Probability of responding to intervention decreases with time since dropout.

Whitehill, Reich, et al. 2015
Whitehill, et al. 2017
Navigating the trade-off

• In Whitehill, Reich, et al. 2015, we explored a simple control-theoretic approach to decide if/when to intervene:

• Essentially, choose threshold on \( P(\text{DropOut} \mid f_w) \) to maximize intervention response, subject to constraint on false alarm rate.
Experiment on 2 MOOCs

• In 2015 we conducted a randomized-controlled trial on 2 HarvardX MOOCs:
  • HLS2x (“ContractsX”)
  • PH525x (“Statistics and R for the Life Sciences”).

• Each of N=8837 participants was randomly assigned to either Control or Intervention condition.

• **Intervention**: send an email to students whom the automatic detector believed had dropped out at each week $w$. 
Results

• Dropped-out students in the Intervention group “came back” into the MOOC stat. sig. more quickly than students in Control group ($p<0.001$).

• One survey respondent wrote:

  \textit{I was not allocating time for edX, but receiving your survey e-mail recaptured my attention.}

• This suggests that even a simple email intervention can affect students’ behavior.
How to measure dropout prediction accuracy?

• In order to facilitate interventions, dropout detectors must be ready before a course starts.

• But the target labels for training+testing become available only at the end of the course.
How to measure dropout prediction accuracy?

• To-date, most research on MOOC drop-out prediction has simply ignored this issue.

• I.e., test accuracy is measured on same course used for training.

• This is tantamount to pretending we could “go back in time” to deploy the detector.
How to measure dropout prediction accuracy?

• Possible solutions:
  • Train on an earlier course.
  • Train on proxy labels that are available earlier in the MOOC.

![Diagram showing the process of training and testing features for dropout prediction.

- **Train:**
  - Features
  - Persist?

- **Test:**
  - Features
  - Certify?

---

Whitehill, Reich, et al. 2015
Whitehill, et al. 2017
How to measure dropout prediction accuracy?

• Training on same course can overestimate accuracy by several %.

• Proxy labels are surprisingly effective.
How to measure dropout prediction accuracy?

- When interpreting accuracy estimates, it is important to consider how the accuracy would impact the effectiveness of an actual intervention.

![Dropout Prediction Accuracy: Comparison across Approaches](Image)
Crowdsourcing 399
Tutorial Videos on Logarithms

Growing $\mathcal{A}$ into $\mathcal{A}' \supseteq \mathcal{A}$
Project goal

• How can we massively expand the action space $\mathcal{A}$ (hints, explanations, practice problems, etc.), that the teacher can choose from?

• Larger $\mathcal{A} \Rightarrow$ potentially higher average learning gains in one-size-fits-all teaching approach.

• Given a large & diverse action space, we can personalize how we teach each student.
Crowdsourcing to facilitate personalized learning

• Diversity of action space can be useful to capitalize on possible interactions between resources and learners:
  
  • **Prior knowledge** that is assumed by the curriculum.
  
  • **Role model effects**, e.g., some women may learn better from female teachers (Dee 2005; Paredes 2014).
  
  • **Language complexity** of the curriculum compared to the language proficiency of the learner (Haag, et al. 2013).
  
  • **Affective state** of the teacher compared to the affective state of the learner.
Research questions

1. Can we design a Human Intelligence Task for Amazon Mechanical Turk (MTurk) so that ordinary people create and share novel video-based learning resources?

2. What kinds of diversity do the resources exhibit?

3. How effective are they for helping students to learn?
Project focus

• In contrast to prior work (Williams, et al. 2016; Chen, et al. 2016), we are interested in collecting video-based (rather than text-based) resources.

• Multimedia videos such as whiteboard animations can help focus students’ attention.

• Multimedia presentations can lead to greater knowledge retention compared to static text-based presentations (Türkay 2016).
Project focus

• We focused on crowdsourcing tutorial videos of how to solve problems about logarithms, e.g.,

  “Solve for \( x \): \( \log_3 x = 27 \).”

• Logarithms:
  
  • Easy enough that many people know what they are.
  
  • Hard enough that many people are confused.
Problems to explain

• We collected tutorials for the following 18 problems:

<table>
<thead>
<tr>
<th>Basic Logarithms</th>
<th>Logarithms and Variables</th>
<th>Equations with Logarithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplify:</td>
<td>Simplify:</td>
<td>Solve:</td>
</tr>
<tr>
<td>log₃ 1 =</td>
<td>log₉ 1 =</td>
<td>log₃(x - 1) = 4</td>
</tr>
<tr>
<td>log 100 =</td>
<td>log₁₅ 125 =</td>
<td>x log₄ 16 = 3</td>
</tr>
<tr>
<td>log₁₀ 1000 =</td>
<td>log₁₆ x² =</td>
<td>z log₁₀ √10 = 4</td>
</tr>
<tr>
<td>log₃ 81 =</td>
<td>log₆ w 1/₁ =</td>
<td>y log₁₀ 1000 = 3</td>
</tr>
<tr>
<td>log₂ 8 =</td>
<td>log₁₁/₄ =</td>
<td></td>
</tr>
</tbody>
</table>

3.1 Correctness

In a pilot run of the experiment, we found that several of the crowdsourced videos were difficult to read. For example, a video that was deemed correct because it began with the problem and ended with the correct solution, and contained a video to be “correct” if it began with the problem explanation how to solve it. Finally, we presented the concrete problem to be solved – e.g., “Simplify equations with Logarithms

Equations with Logarithms

Solve:

log₃(x - 1) = 4
z log₁₀ √10 = 4

3.2 Format

We observed two general approaches that teachers used to transform a problem step-by-step into the solution. For instance, one teacher derived the solutions to the problems. In some explanations, we noted that log₂ x" – and asked them to create and upload a video explanation. Finally, we can use the fact that the power to which 10 must be raised to equal x to derive 4 log₂ 1000 = 3. In other explanations, the teacher emphasized the need for both a large, diverse set of explanations as well as widely used in India [6]. This also highlights the variability in the geographical origin and dialect of the speaker.

3.3 Handwriting

As shown in Figure 3, there was diversity in the presentation of handwriting in subsequently submitted videos. The five videos shown in the figure illustrate the most common styles; these include a static Powerpoint slide to which the instructor points (5), a step-by-step “Powerpoint”-style presentation; (4) a step-by-step computer screen; (3) speaking directly to the learner in a face video along with written materials to show the derivation itself was correct. For example, one teacher would note that log₃ 1 = 0, log₁₀ 100 = 2, log₁₀ 1000 = 3, log₉ 1 = 0, log₁₅ 125 = 3, log₁₆ x² = log₁₆ x² = log₆ w 1/₁ = log₆ w 1/₁ = log₁₁/₄ = log₁₁/₄ = log₃(x - 1) = 4, x log₄ 16 = 3, z log₁₀ √10 = 4, y log₁₀ 1000 = 3.
MTurk Human Intelligence Task (HIT)

• The HIT that we posted to MTurk contained:

  • **Acceptance criteria** that the resource be a novel, mathematically correct, video tutorial.

  • **Examples** of good tutorials from YouTube and Khan Academy.

  • **Suggestions** to improve quality, e.g., examples of good versus bad handwriting.
HIT on MTurk:

Consent Form & Video Recording Release Form

... You will then be asked to create a novel video in which you explain how to solve a short mathematical exercise: <PROBLEM>. The content and format of the video are up to you, but the video must address the problem and must be mathematically correct. For example, the video might contain a screenshot showing an electronic “blackboard” on which you explain how to answer the problem. Alternatively, you might prefer to talk into a web camera and record a video of your face and your voice...

Survey

Please answer the questions below. When you are done, click "Next".

1. How old are you (in years)?
2. What is your gender?
3. What is the highest level of education you have completed? …
4. How much do you enjoy mathematics? …
5. How do you prefer to learn something new? …

Sample Problems & Explanations

This page contains some example videos that explain how to solve math problems. Please watch the videos carefully so you know what we are looking for in this HIT.

Hints on Making a Good Video

When you make your video, you may sometimes record images of your own handwriting. Please look at the following handwriting examples so you know what distinguishes a good video from a bad video. Note that a bad video may be rejected due to poor image quality.

The following 3 examples are OK – the writing is dark, big, and clear.

The following 3 examples are not OK – the writing is too small, blurry, and/or hard to read.

Problem & Instructions

Please examine the following math problem: <PROBLEM>

Instructions:

1. Think carefully about how you would explain to someone else how to solve this problem.
2. Create a video that explains how to solve the problem.
3. Upload the video to our server.

Rules:

- Your video must explain how to answer the following math problem: <PROBLEM>
- Your video must be original - it cannot be an existing video.
- Your video must be mathematically correct.
- Your video may not contain any images of children (<18 years old).
- Your video may not contain any nudity or profanity.

Submission

...
Collection results

- Over 2 periods of 2 weeks each, we collected unique **399 videos** from **66 different teachers**.
  - 17% authors were female, 18-55 years old (self-report)
  - Most videos **1-3 minutes** long.
  - Each teacher could contribute 1 video for each topic (up to 18 total topics), for **$5/video**.
  - Total cost: $2800 (including 40% commission to Amazon).
Correctness

• Videos were labeled for correctness:

  • **Fully correct**: no false math statements (note we did not attempt to judge pedagogical quality).

  • **Incorrect**: at least 1 false math statement.

  • **Borderline**: minor issues such as calling an expression an “equation”.

  • **Improper submission**: not novel, not a video (e.g., just a still image).
Correctness

• Results:

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully correct</td>
<td>81%</td>
</tr>
<tr>
<td>Incorrect</td>
<td>11%</td>
</tr>
<tr>
<td>Borderline</td>
<td>5%</td>
</tr>
<tr>
<td>Improper submission</td>
<td>3%</td>
</tr>
</tbody>
</table>
Sample (correct)

Video link
Sample (incorrect)

\[
\frac{\log \frac{1}{4}}{\log \frac{1}{2}} = \frac{\log \frac{1}{4}}{\log \frac{1}{2}} = \frac{1}{2}
\]
Diversity

• We qualitatively examined the diversity of submissions along several dimensions including:
  • Presentation style
  • Pedagogical path
  • Language differences (all in English)
Presentation style

Video 1

Video 2

Video 3

Video 4

Video 5

Whitehill & Seltzer 2017
"Solve for $x$: $x \log_4 16 = 3$"
Diversity of pedagogy

“Solve for $x$: $x \log_4 16 = 3$”

Each of the 17 pedagogical paths was unique.
Diversity of pedagogy

- Multiple degrees of granularity.
- Some learners may prefer a slower/faster approach.

Whitehill & Seltzer 2017
Diversity of language

• Many videos contained phrasing such as “5 into \(y\)”: 
  
  • In North-American English, this phrasing can be confusing because it suggests \(y\) is divided by 5.
  
  • In India (and other countries), this expresses 5 times \(y\).
Effectiveness for learning

• Do the crowdsourced videos actually help students to learn?

• We don’t expect all the videos to be good, but hopefully some of them are.

• Looking for “diamonds in the rough”.
Experiment 1

• Randomly pick 40 videos that were verified as fully correct.

• Also pick a “control” video on math topic unrelated to logarithms.

• Does watching one of the 40 crowdsourced videos help students to learn better than the control video?
Experiment 1

- **Participants**: $N=200$ users on MTurk (separate experiment from crowdsourcing task).

- **Protocol**:
  1. Pretest on logarithms.
  2. Watch randomly assigned video (control: 20%; one of the 40 videos: 80% (2% each)).
  3. Posttest on logarithms.

- **Dependent variable**: Average learning gains $G_k$ (Posttest minus Pretest) for each video $k$. 
Experiment 1: Results

- Average $G_k$ (10.5%) over all 40 crowdsourced videos was stat. sig. ($p<0.001$) higher than for control video (4.5%).

- A few videos delivered much better performance.
Experiment 2

• How good were the best 4 videos from Experiment 1 compared to a similar video on logarithms from an “expert teacher”?

• We chose a Khan Academy video for comparison:

  • 924,520 views as of October 20, 2016.
  
  • 7 minutes long (3x longer than crowdsourced videos).
  
  • Broad tutorial on logarithm (not just about 1 specific problem).
Experiment 2

• **Participants**: $N=250$ users on MTurk (separate experiment from crowdsourcing task).

• **Protocol**:
  1. Pretest on logarithms.
  2. Watch randomly assigned video (Khan: 20%; one of the 4 videos: 80% (20% each)).
  3. Posttest on logarithms.

• **Dependent variable**: Average learning gains $G_k$ (Posttest minus Pretest) for each video $k$. 
Experiment 2: Results

<table>
<thead>
<tr>
<th>Video</th>
<th>Participants</th>
<th>$G_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58</td>
<td>0.1416</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>0.1140</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>0.0942</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>0.0932</td>
</tr>
<tr>
<td>Khan</td>
<td>58</td>
<td>0.1506</td>
</tr>
</tbody>
</table>

- Average learning gains of best crowdsourced video was 14.16% (versus 15.06% for Khan). Difference was not stat. sig. ($p=0.82$).

- Result suggests that crowdsourcing can uncover some learning resources that work very well for students on average.

- Also possible that higher learning gains can be attained by personalizing which student receives which video.
Future challenges
Learning from data (?)

• The promise of “big data” in education is that we will learn how humans learn and thereby teach more effectively.

• Has this happened yet?

• As Justin Reich pointed out (Science, 2015), we don’t need massive clickstream logs to know that “there is a positive correlation between student activity and success”.
Learning from data (?)

- One big impediment to learning from MOOC and ITS datasets is that they are mostly observational rather than experimental.

- **Observational** data: students decide the actions themselves (e.g., enroll in next course).

- **Experimental** data: actions are assigned randomly (e.g., student X is assigned to teacher Y).
Learning from data (?)

- With observational data, causal inferences of what are the most effective interventions can be confounded.

- **Confound**: (often unobserved) variable that predicts both the intervention and the outcome, e.g.:
  - Students who do X perform better than students who do Y.
  - So is X better than Y? Not necessarily.
  - Maybe the more motivated students tend to do X and perform better? (Motivation is a potential confound.)
Learning from data (?)

• There are sometimes ways of eliminating confounds, by using statistical matching methods or regression discontinuities.

• But randomized-controlled trials (RCTs) are still very important (c.f. Justin Reich’s talk) to estimate the effects of actions on learners: $P(s_{t+1} \mid s_t, a_t)$
Personalized learning

• While intuitively appealing, the empirical evidence in support of personalized learning software is surprisingly limited (Yarnall, et al. 2016).

• Identifying and measuring the key interactions between learning resources and learners will be important to realize more effective personalized learning systems.
Privacy

• Especially when considering recording & analyzing students’ face data, privacy becomes a big concern.

• Rather than transmit the data, perhaps the facial signals could be measured and used locally?

• Important to make a strong empirical case that using learners’ data actually improves their personal learning.
Thank you, TU Delft, for the opportunity to speak.
Questions