Machine-learned Computational models to Assess Ill-defined Constructs

Sidney K. D’Mello
University of Colorado Boulder

sidney.dmello@colorado.edu
www.colorado.edu/ics/sidney-dmello
August 23, 2019
• traditional methods (classical test theory, item response theory, evidence centered design) have been invaluable for assessing a range of constructs (e.g., knowledge, skills)
• but what about “ill-defined” constructs that cannot be precisely defined, are ephemeral states, especially in situ?

• machine-learned, computational models are essential
  • when constructs are “ill-defined” like emotion, collaboration
  • when there are no adequate theoretical mechanistic accounts
  • when underlying models are “multilevel circular causal”

• models can promote change via intervention or reflection
• the art lies in how they are constructed and evaluated
• and in setting realistic expectations and contexts of use
conceptual model [affect example]
D’Mello, Kappas, & Gratch (2018)
If a researcher wanted to examine the effect of a diet pill on weight loss she might give some participants the diet pill and other participants would receive a sugar pill that looked identical to it.
ubiquity of mind wandering
method (Faber, Bixler, & D’Mello, 2018)
- model estimates correlated with self-reported mind wandering ($r = .400$)
- correlated with comprehension ($r = -.374$) stronger than self-reports ($r = -.208$)
- models robust to missing data and internally consistent ($r = .751$)
- page-level predictions moderate — precision of 72.2%; recall of 67.4%
- fewer but longer fixations and fewer horizontal saccades related to mind wandering
real-time intervention (Mills, et al., in review)
method

- 70 participants read a book on surface tension in liquids
- randomly assigned to intervention or yoked-control
- tested for text- AND inference- level comprehension after reading AND one week later (parallel forms)
out of the lab and into the wild (Hutt et al., 2019)
• tracking validity between 75% (both eyes) and 95% (one eye)
• moderately accurate at mind wandering detection (precision .55; recall .65)
• model predictions correlated with learning ($r = -20$)
can eye movements predict comprehension?
very accurate for textbase-level comprehension assessed during reading (AUROC = 0.9; r = 0.68) Gregg & D’Mello (in review)
motivation

- surprising lack of consistency in literature
- very little research on long connected texts, especially after reading
- tested weak vs. strong association hypotheses ($R^2$ of 1% vs. 10%)

methods

- datasets 1 and 2: predict textbase-level comprehension 30-mins after reading one long connected text
- dataset 3: predict textbase- and inference-level comprehension after reading up to 8 short texts
- focused on seven eye gaze features and reading times
- simple cross-validated regression models

what about comprehension after reading?
(Gregg, Bixler, & D’Mello, in review)
• moderate cross-validated correlations between observed and predicted comprehension
• models from one study generalized to another
• more, but shorter, fixations predicted comprehension
• results hold after accounting for mind wandering and exposure to print (author recognition test)
machine-learned computational models of eye movements can assess reading processes and outcomes & can drive intervention
video-based modeling of affect and attention
does frame-of-reference coding help

D’Mello (2016)
it depends…

between judges

agreement with self

it depends…
Physics Playground

online observations

results (AUC)

facial features + body movements

modeling affect from video
(Bosch, et al., 2016)
video-based mind wandering detection

Bosch & D’Mello (2019)
video-based models can provide human-comparable results for affect and attention
speech and language processing for discourse analysis
The districts we studied spend an average of nearly $18,000 per teacher, per year on professional development.

Give teachers a clear, deep understanding of their own performance and progress.
Which of these would you consider authentic?

Teacher: “How does a person become a noble?”
Student: “They’re born into it”
Teacher: “They’re born into it, right? It’s by family. It gets passed down ….

Teacher: “How did that make you guys feel, I mean what was your gut reaction to all that?”
Student: “Ashamed”
Teacher: “Ashamed in what way?”

authentic questions
128 hours of audio from 132 classes by 14 teachers from 7 schools

teacher mic
(Samson Airline 300)

Classroom mic

mixer
(M-Audio M-Track)
Speech and Language Processing

Data collection (132 observations from 27 classes by 14 teachers in 7 schools)

Teacher mic (Samson Airline 77)

Teacher audio (7,663 mins total)

Automatic speech segmentation (45k utterances)

Bing speech recognition (text transcript)

Live coding and offline refinement of codes

Gold-standard authentic question codes

Data for machine learning

Natural language processing (word, sentence, and discourse level features)

Machine Learning & Validation

Leave-one-teacher-out cross validation

Learn M5P regression trees for k-1 teachers (in gray)

Apply model to generate estimates for held-out teacher (in black)

Repeat until each teacher is held out once

Pool computer-estimates and compare with gold-standard codes
computer scores of authenticity correlated with human codes ($r = .686$)
<table>
<thead>
<tr>
<th>Streamlined recording</th>
<th>Teachers collect data</th>
<th>Utterance-level coding</th>
<th>Spreadsheet-based coding</th>
<th>Expanded codes</th>
<th>Multilevel modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher mic only</td>
<td>127 class sessions from 16 teachers in western Penn.</td>
<td>Cloud-based automatic speech recognition generates utterances</td>
<td>Direct coding of utterances into using excel macros and foot pedals for audio</td>
<td>Expansive set of codes including teacher-led and transactional discourse</td>
<td>Modeling at teacher, class, and utterance levels</td>
</tr>
<tr>
<td>Moving to lapel/smartphone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

new approach
Positive and negative feedback lets students know if their comments are on the right track or if they need to adjust. Both positive and negative feedback can promote learning, but experts suggest a 6:1 ratio of positive to negative.

### Design of Feedback App

**My Notes:**
Students were engaged, but had some issues grasping concepts.

- **Instructional Talk**
- **Question Asking**

![Graph showing trends in feedback]

- **Trends**
  - 2 Selected
  - Subjective Questions
  - Amer. Lit. Period 2
  - Amer. Lit. Period 3
  - Recommended
models of spoken language can capture complex aspects of discourse in noisy environments
description
• funded by Intelligence Advanced Research Projects Activity (IARPA)
• challenge was to robustly predict psychological traits, health/wellbeing, and job performance in the real-world from sensors alone?
• target correlation of 0.5 on a blinded sample
• do it all in 16 months

our approach
• Project Tessarae - 10 PIs from 8 universities
• collected data from 757 US information workers for 1-year
• four sensors (wearable, Beacons, phone agent, social media)

results
• modeling social, lifestyle, tech use, physiology/behavior, & context
• ensemble-based machine learning approach for robustness
• average correlation of 0.21 [0.08 to 0.41] on 14 constructs
patterns of life

These results are based on an average of 63 weeks of data.
machine-learned, computational models can enhance assessment
• machine-learning when theory/mechanisms are sparse
• data is abundant and sufficiently complex (nonlinearities)
• models can promote change with intervention and/or reflection

tips on constructing models
• reliance on theory without being overly constrained by it
• striving for parsimony rather than chasing fads (deep learning)
• important to go beyond minimizing validation loss
• explainability, real-time applicability, fairness, & generalizability
things to consider when assessing ill-defined constructs

- defining constructs – don’t really need precise definitions
- reliability concerns – reliability important but not a show stopper
- quantify performance – external sources critical
- what is good performance? – beyond chance probabilistic
- how good is good enough? – good for what purpose?

looking into the future

- standardized testing
- game-based assessments & performance tasks
- machine-learned computational models for specific tasks
- is the future robust multimodal sensing in context?

concluding thoughts
www.colorado.edu/ics/sidney-dmello
sidney.dmello@colorado.edu