SIOP 2023 ML Competition

Isaac Thompson, Georgi Yankov, Sebastian Marin & Nick Koenig

SIOP ANNUAL CONFERENCE
BOSTON and ONLINE • April 19–22, 2023

Can I-O Psychology Keep Up?

Methodology example:

- **2013** word2vec
- **2014** seq2seq
- **2016** bi-directional LSTMs
- **2017** Transformers
- **2018** BERT
- **2019** RoBERTa
- **2020** GPT3 (175 billion parameters)
- **2023** ChatGPT

Publishing example:

- **2017** deep learning invention to score text (LSTM), takes a year or so to write a paper
- **2019** While in 2 year R&R, transformers make that approach obsolete
- **2020** Add transformers, R&R
- **2021** Rejection and on to new journal,
- **2022** accepted final edits
- **2023** in print; ChatGPT comes out

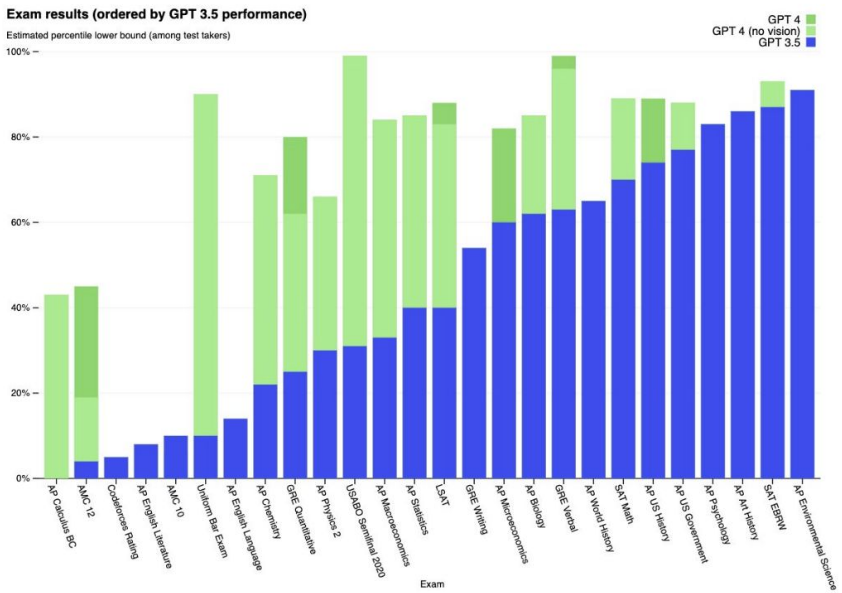
Can I-O Psychology Keep Up?

Used to joke that IO will fall behind; not as much of a joke now.

- Zone 5 occupations (those with this highest level of educational requirements) will be the most disrupted by LLMs
- One study puts us in the top 20% of occupations; another study in the top 7% (57th out of 774).

What can AI not do? What can it do?

- Assessments:
  - Write items
  - Respond to those items
  - Rate those items (as good as MTURK)
  - Automate scoring of those items
  - Explain the rating
  - Generalize to new items
- Can it do a lit review (autoGPT)?
- Can it generate code?
- Can it pass comps?
- How fast is it changing now?



- **Few month difference from ChatGPT 3.5 to 4.0.**
- **Notable examples:** Bar exam from 10% to 90%; Easy coding exam (leetcode) from 29% to 75%; Quant GRE 25th to 88th.

What can we do?

- We need open data, open (reproducible) code, collaboration at scale, living benchmarks

ENTER ML COMPETITION

What is a Machine Learning competition?

A data set is released (training set) with a problem statement

Community attempts to solve the problem statement, empirically

Scaled evaluation of approaches is accomplished via an online portal where predictions on a private data set (dev set; public leaderboard) are assessed empirically and automatically

Best generalizable solution wins as teams submit to a final private leaderboard that no one sees on a third data set (test set)

Winners are decided based on the empirical quality of their work

The benchmark lives beyond the competition as new methods become available

How we do it @ SIOP

Data sponsor to open source I-O data; paired with knowledge of what are hot problems facing the field; and an evaluation schema is created to rank teams.

Open registration to anyone and everyone (272 individual emails registered this year)

Scaled evaluation (via eval.ai); we codify that ranking schema. Teams submit their predictions. 28 teams made it to the leaderboard (average 4.5 participants per team in the past). Over 1,200 unique prediction sets submitted.

Announce winners (today)

Put on Github: all the data, winning solutions end to end code bases, & presentations.

History and Purpose of SIOP's ML Competitions

1. **2018**: Predict turnover: Eli Lilly and Company
2. **2019**: Predict self-report personality from open ended text: Shaker/Modern Hire
3. **2020-2021**: Predict who to hire; balancing fairness and validity: Walmart
4. **2023**: Predict assessment center ratings of decision making from open ended text: DDI

HUGE THANKS TO DDI for this living data set.

Let's dive into the data.

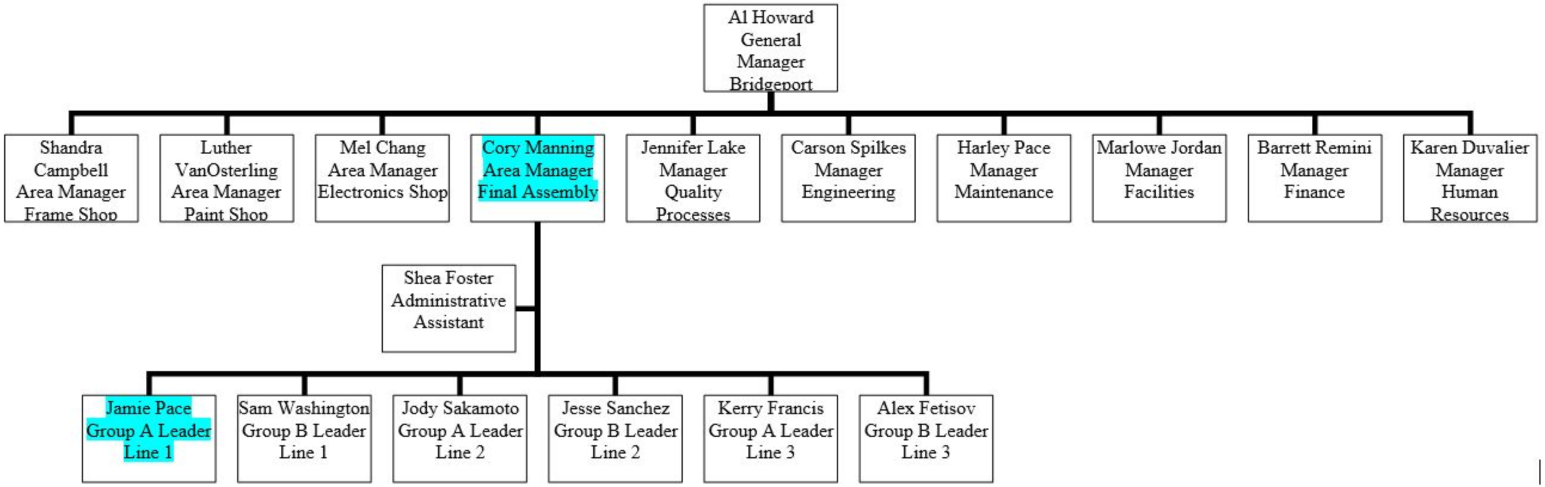
Competition Data

Three archival assessment center platforms (early 2010s) mixed in, and rationale for doing that.

- 61% manufacturing leader level I
- 22% manufacturing leader level II (boss of level I)
- 17% service leader

Immersive experience and a background storyline in each

Competition was only on the operational challenges (in-basket emails)



Assessor Scoring

- One assessor scores all exercises and the scoring is left unchallenged in the integration.
- For some *exercises X behaviors* the scoring rubric specifies exactly what is expected (e.g., Interprets Information in Exercise 3).
- Some exercises have higher priority.
- Some exercises are left uncompleted because of time constraints and personal choices.
- Some behaviors are more essential but we wanted good models for all, thus weighting was equal.
  - In reality, dimension score was a judgement based on rubric of possible behavior score combinations.

Dimension Key Action	Exercise 1	Exercise 2	Exercise 3	Exercise …	Exercise N	Dimension Key Action	Ratings	Rationale
Decision Making						**Decision Making**	**1 2 3 4 5**	
± Identifies Issues		•			•	± Identifies Issues	-- - + ++	
± Gathers information				•		± Gathers Information	-- - + ++	
⇒ Interprets Info			•			⇒ Interprets Info	-- - + ++	
⇒ Chooses Appropriate Action	•		•		•	⇒ Chooses Appr Action	-- - + ++	
± Commits to Action	•		•		•	± Commits to Action	-- - + ++	
± Involves Others	•					± Involves Others	-- - + ++	

Challenges and Opportunities for Automated AC Scoring

The competition data is scored with an older approach we no longer use operationally.

- Feedback reporting scale: Need for development (1-2) - Proficient (3-5) - Strong (6-7)

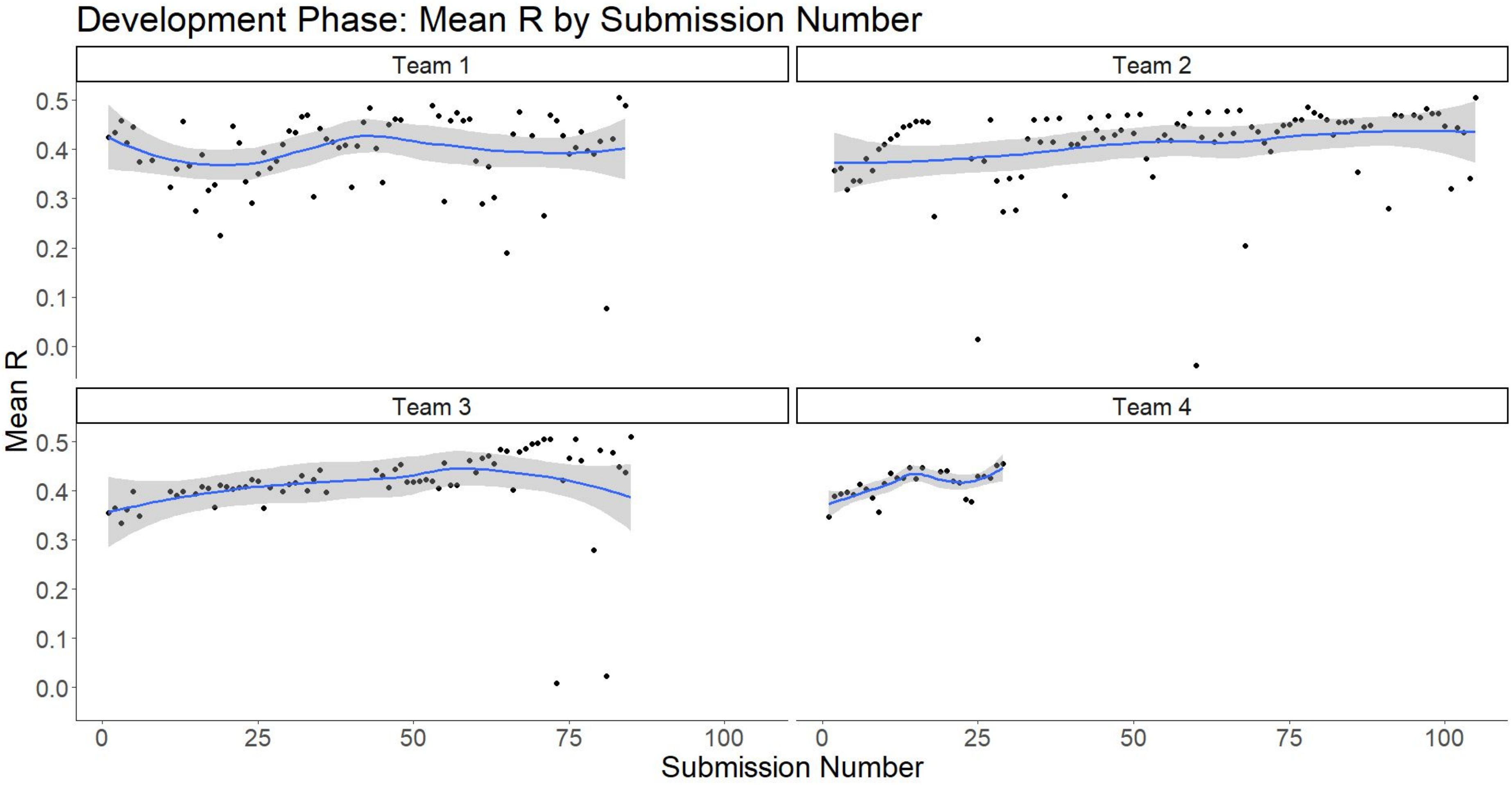
Challenge 1: Long, open-ended texts, sometimes full of typos and mistakes

Challenge 2: Behavior appearing anywhere in text, sometimes with idiosyncratic language

Challenge 3: Training data is not in the tens of thousands

Opportunities: Transformers-based models, multiclass models, ensembles, and …….?

Winner Announcement 🎉



Development Phase: Mean R by Submission Number
Team 1
Team 2
Team 3
Team 4
Mean R
0.5
0.4
0.3
0.2
0.1
0.0
0
25
50
75
100
Submission Number

Team	Chooses Appropriate Action	Commits to Action	Gathers Info.	Identifies Issues and Opportunity	Interprets Info.	Involves Others	Decision Making Score	Mean *R*
Team 1	.475	.421	**.430**	**.393**	.507	.340	**.657**	**.520**
Team 2	.478	**.439**	.386	.355	**.518**	**.394**	.609	.501
Team 3	**.500**	.434	.322	.345	.490	.348	.639	.500
Team 4	.496	.425	.354	.353	.490	.327	.609	.488

Team	**Chooses Appropriate Action**	**Commits to Action**	**Gathers Info.**	**Identifies Issues and Opportunity**	**Interprets Info.**	**Involves Others**	**Decision Making Score**	**Mean *R***
Team 1	.475	.421	**.430**	**.393**	.507	.340	**.657**	***.520***
Team 2	.478	**.439**	.386	.355	**.518**	**.394**	.609	.501
Team 3	**.500**	.434	.322	.345	.490	.348	.639	.500
Team 4	.496	.425	.354	.353	.490	.327	.609	.488

↓

Team 0	.500	.439	.43	.393	.518	.394	.657	***.530***

Drum roll 🥁

🥇 Sentient Sentence Sense-AIs

Ivan Hernandez (Virginia Tech), Andrew Cutler (Freelance), Joe Meyer (Erudit), Wewein Nie (Hogan Assessments)

🥈 team__mifflin__

Ammar Ansari (California Baptist University)

🥉 mustafaakben

Mustafa Akben (Elon University)

🏅 GHAAS (Global Hiring at Amazon)

Yizhen Egyn Zhu (Amazon), Dawn Sepehr (Amazon)

Presentations

Discussion

Questions for the Winners

Rapid Fire:

How did you do it, what was your secret sauce?

What would you have done differently?

Where do you see these methods being applied in I-O?

What most impressed you most about the other teams' approaches?

What is your takeaway from participating and winning a ML competition?

What would you like to see in future I-O ML competitions?

Questions from the participants/audience