

Trends on Returning Contestants and Geography at the International Olympiad in Informatics

Authors: E.Lee, T.Reizin, F.E.Wu, F.E.Wu
Presenter: G.F.Ortega

IOI Conference, September 2024

- Analysis on historical IOI data aimed at quantifying intuition about “IOI meta”.
- Two focus topics for the analysis:
 - Returning contestants and their progress over time.
 - Factors that affect country performance, geographic trends.
- Primarily produced data visualizations for exploratory data analysis.
- Goal is to broadly open the conversation on data-driven approaches in training.

Both score and percentile have drawbacks as performance measures, thus we calculate a "scaled score" that is a piecewise linear interpolation.

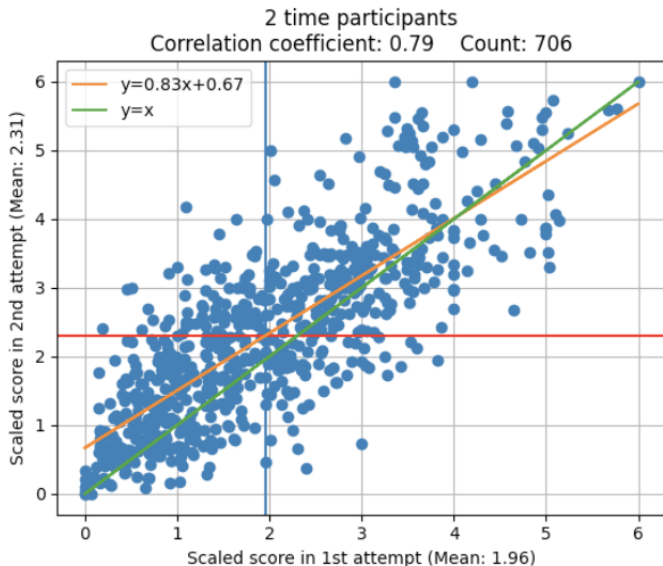
- The lowest score is mapped to 0 and the highest to 6.
- Bronze, silver, and gold cutoffs are mapped to 2, 3, and 4, respectively.
- 25th percentile (middle of no medal group) is mapped to 1, and 96.67th percentile (middle of gold) is mapped to 5.

Returning Contestants Analysis

In this section, we present the analysis of returning contestants at the IOI, along two aspects: returning contestant's performance and improvement across multiple IOIs, and progress of contestants with multiple IOIs.

- In total, there are 1038 participants with more than one IOI from 2011 to 2023, which provides ample data for the analysis in this section.
- The trends shown are mostly rather strong and unlikely to be affected by random noise in the data.

Cross-distribution of Performance Across Multiple IOIs



Cross-distribution of Performance Across Multiple IOIs

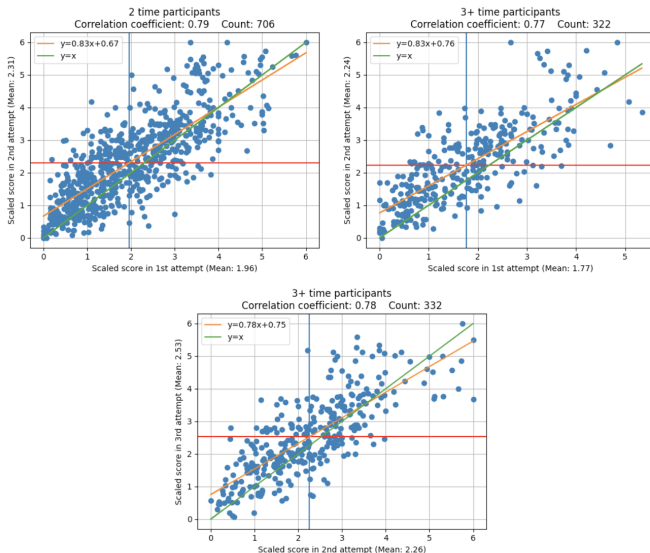


Fig. 1. Scatterplot of returning contestants' performance in consecutive IOIs.

Cross-distribution of Performance Across Multiple IOIs

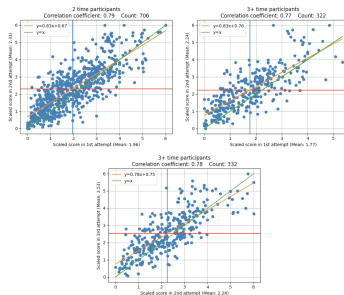
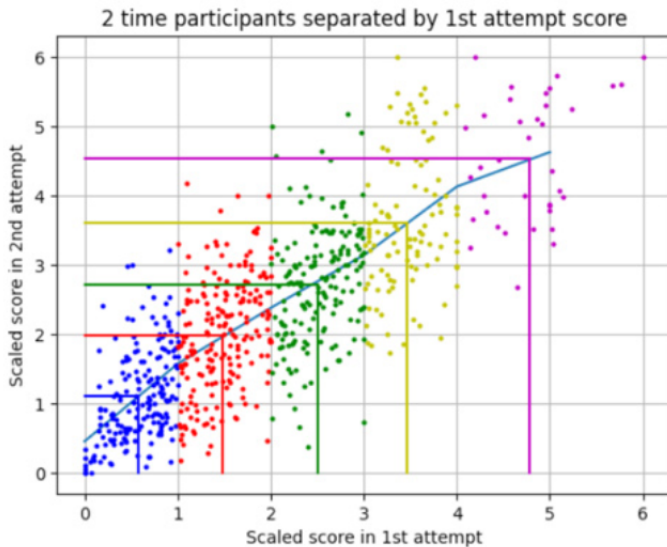


Fig. 1. Scatterplot of returning contestants' performance in consecutive IOIs.

- The correlation coefficients are calculated to range from 0.77 to 0.79.
- The regression coefficients are shown to range from 0.78 to 0.83, again indicating strong dependence across IOIs participated.
- The regression line, for most of the corresponding datapoint x-values, is above the 45-degree line, and most of the datapoints in the scatterplot are above the 45-degree line.

Regression spline of scaled scores in consecutive IOI's



Regression spline of scaled scores in consecutive IOI's

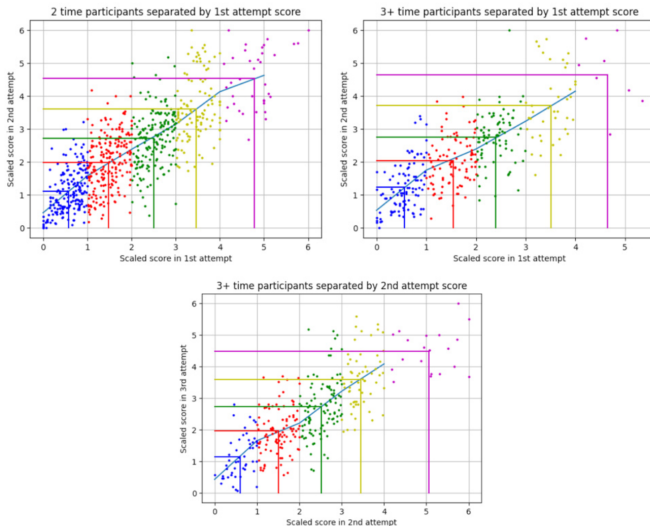


Fig. 2: Regression splines of scaled scores in consecutive IOI attempts

Regression spline of scaled scores in consecutive IOI's

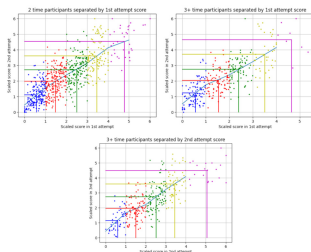


Fig. 2: Regression splines of scaled scores in consecutive IOI attempts

- We plot a linear spline regression with knots at integer scaled scores since it's unclear whether a single trend is appropriate for all levels.
- From Fig. 2, a positive slope within each scaled score interval is consistently seen. The slopes, however, vary substantially.
- The slope in the $[0, 1)$ scaled score interval is substantially larger. This observation may be interpreted as that having a baseline level is crucial towards improving further towards the IOI medal level.

Cross-distribution of Performance Across Multiple IOIs

- Those performing below the 25th percentile, are very unlikely (under 10% chance) to win a medal at the following IOI.
- Those performing above the 25th percentile, have roughly even odds to do so.
- Around 40% of bronze medalists achieve a silver medal or better at the next IOI (under 5% get gold).
- Among silver medalists, around 35% manage to improve to a gold medal at the following IOI, while around 20% drop to bronze.

Quantile plot of improvement, by initial result bracket

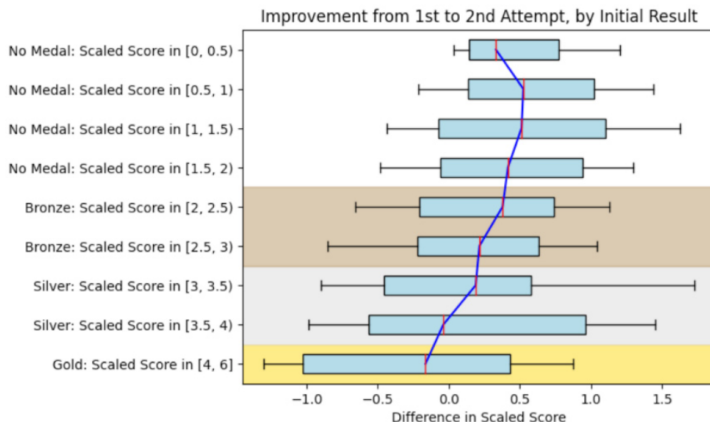


Fig. 3. Quantile plot of improvement across consecutive IOIs, separated by initial result bucket.

Quantile plot of improvement, by initial result bracket

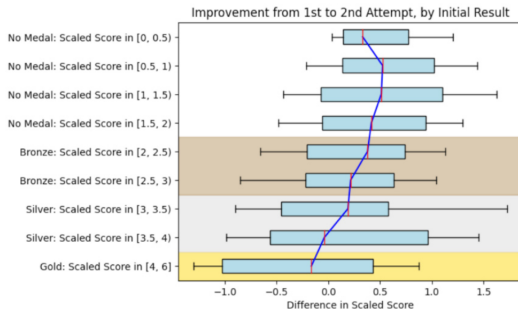


Fig. 3. Quantile plot of improvement across consecutive IOIs, separated by initial result bucket.

- The plot shows a general trend of the average improvement decreasing as the initial IOI performance increases.
- The variance in improvement tends to be large, with the 10th and 90th percentiles spanning more than one (often two) scaled score point.

Progression of mean scaled score in consecutive IOIs

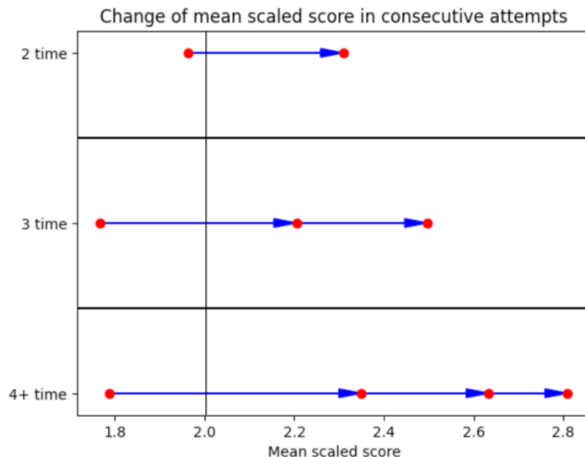


Fig. 5. Progression of mean scaled score in consecutive IOIs, separated by number of IOIs.

Progress of Multiple-IOI Contestants Over Time

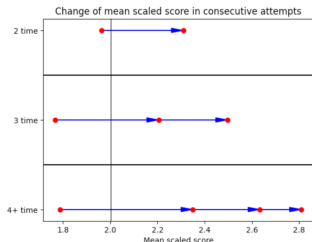


Fig. 5. Progression of mean scaled score in consecutive IOIs, separated by number of IOIs.

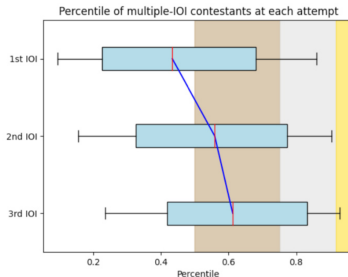


Fig. 6. Quantile plot of percentile distribution separated by ordering of IOI attempt.

- The contestants tend to improve over IOIs, though the improvement is typically small. Improvements tend to decrease over time.
- Contestants with more IOIs tend to improve faster, while starting at a lower level and reaching a higher level.
- Fig. 6 shows that multiple IOI contestants start out slightly worse than the average contestant, they eventually improve to be stronger.

Performance change in 1st to 2nd, vs 2nd to 3rd IOI

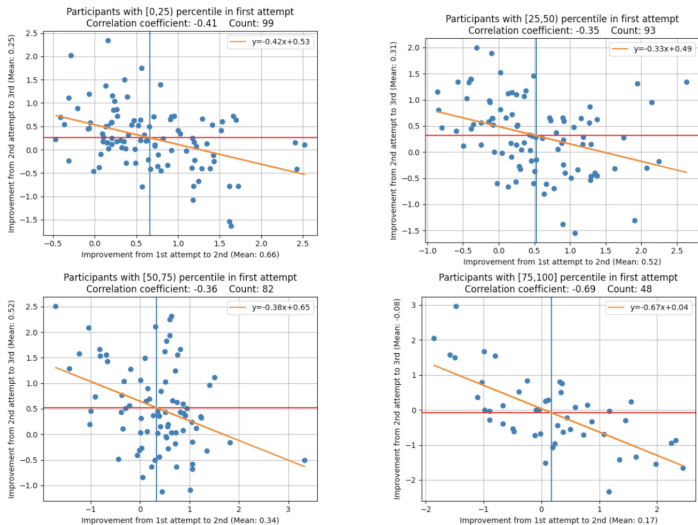


Fig. 8. Scatterplot of performance changes in consecutive IOIs, separated by initial performance.

Performance change in 1st to 2nd, vs 2nd to 3rd IOI

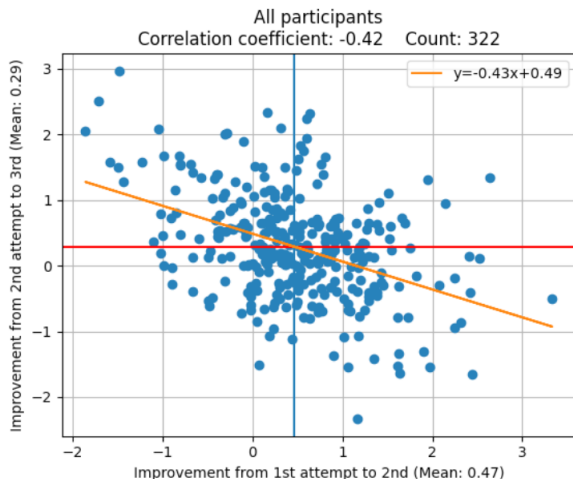


Fig. 7. Scatterplot of performance changes in consecutive IOIs, aggregated data.

Progress of Multiple-IOI Contestants Over Time

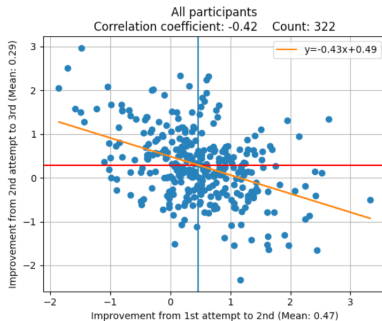


Fig. 7. Scatterplot of performance changes in consecutive IOIs, aggregated data.

- The plot shows the “regression to the mean” phenomenon, indicating that there is a substantial amount of idiosyncratic variance in IOI performance at all levels of competition.

Country Classification and Analysis

- What statistics of countries best explain the general strength of that country's performance at the IOI?
- How well do these statistics explain the country's strength?
- Countries with at least 26 contestants from IOI 2011 to 2023 were included in the analysis.
- A country's performance is measured as the average percentile of its team members throughout its years of participation.

Regional categories of countries

Countries are grouped into 11 categories based on the United Nations geoscheme. A few modifications were made to balance category sizes.

- Asia
 - ① Southeastern Asia
 - ② Eastern Asia
 - ③ Central and Southern Asia
 - ④ Western Asia
- Europe
 - ⑤ Eastern Europe
 - ⑥ Southern Europe
 - ⑦ Northern Europe
 - ⑧ Western Europe
- Africa
 - ⑨ Africa
- Latin America
 - ⑩ Latin America
- Anglosphere
 - ⑪ Anglosphere

Population and Human Development Index

- An initial analysis was done analyzing the correlations between the population and HDI to each country's performance.
- A larger population would be expected to correlate positively with performance due to having a larger talent pool to draw from.
- Meanwhile, a greater HDI indicates a greater availability of resources used to nurture and train talented individuals.

Avg. IOI percentile by country, vs log Population and HDI

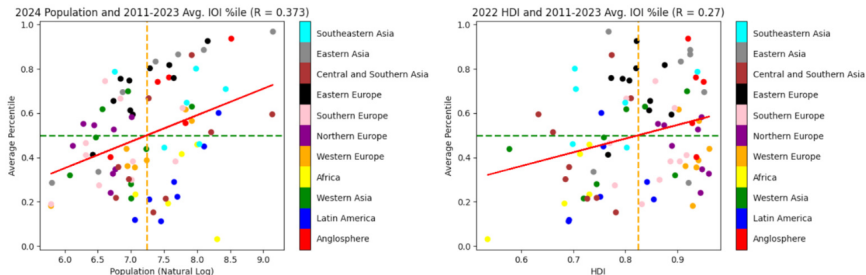


Fig. 9. Scatterplot of average IOI percentile by country, with log Population and HDI.

- Both of these correlations are weak but positive.
- Not a lot of the variation is explained solely by these two variables.

Regression of Avg. IOI %ile vs. log population and HDI

Table 3
Regression output of average IOI percentile on HDI and log population

OLS Regression Results						
=====						
Dep. Variable:	Average Percentile	R-squared:	0.312			
Model:	OLS	Adj. R-squared:	0.294			
Method:	Least Squares	F-statistic:	17.88			
No. Observations:	82	AIC:	-35.68			
Df Residuals:	79	BIC:	-28.46			
Df Model:	2					
=====						
	coef	std err	t	P> t	[0.025	0.975]

constant	-1.5287	0.340	-4.495	0.000	-2.206	-0.852
2022 HDI	1.0051	0.226	4.451	0.000	0.556	1.455
2024 Population (ln)	0.1657	0.032	5.237	0.000	0.103	0.229

- The regression output corroborates the fact that much of the variation is unexplained, yet the analysis is statistically significant.

Summary of residuals by geographic region

Table 4
Residual summary statistics by geographical region

Region	Residual Mean	Residual Stdev.	Count
Eastern Europe	0.220	0.094	10
Eastern Asia	0.122	0.106	7
Southeastern Asia	0.105	0.153	6
Western Asia	0.059	0.151	8
Anglosphere	0.007	0.131	5
Southern Europe	-0.009	0.195	10
Central & Southern Asia	-0.018	0.193	9
Northern Europe	-0.044	0.173	8
Africa	-0.170	0.120	5
Western Europe	-0.171	0.057	7
Latin America	-0.221	0.115	7

- An analysis by region shows that the regions of Eastern Europe, Eastern Asia, and Southeastern Asia have the highest performances relative to their population and HDI statistics.
- This aligns with the perceived emphasis placed on scientific education in these regions.

CodeForces participant count vs country average %ile

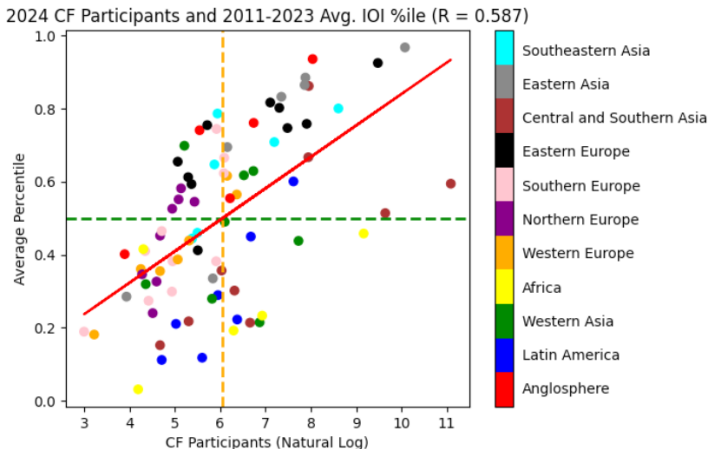


Fig. 10. Scatterplot of CodeForces participant count and country average percentile.

CodeForces participant count vs country average %ile

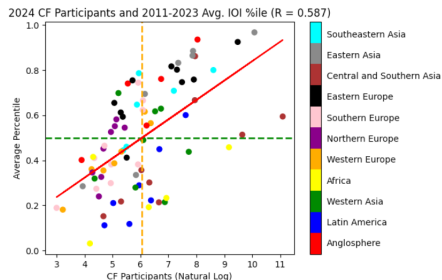


Fig. 10. Scatterplot of CodeForces participant count and country average percentile.

- The number of registered Codeforces participants was used to measure approximate levels of interest of competitive programming in each country.
- A strong, positive correlation is present, showing that general interest is a major factor of IOI performance.

Adding CF density to regression

Table 5
Regression output of average IOI percentile on HDI, log population, and log CF density

OLS Regression Results						
=====						
Dep. Variable:	Average Percentile	R-squared:	0.602			
Model:	OLS	Adj. R-squared:	0.587			
Method:	Least Squares	F-statistic:	39.32			
No. Observations:	82	AIC:	-78.59			
Df Residuals:	78	BIC:	-68.96			
Df Model:	3					
=====						
	coef	stderr	t	P> t	[0.025	0.975]

constant	-1.1855	0.264	-4.487	0.000	-1.711	-0.660
2022 HDI	1.2334	0.175	7.030	0.000	0.884	1.583
2024 Population (ln)	0.2608	0.027	9.553	0.000	0.206	0.315
2024 CF Density (ln)	0.2645	0.035	7.542	0.000	0.195	0.334

- Adding the CF density (number of registered users over country's population) explains much more of the variance seen in countries' performances. R^2 is 0.601 vs. 0.312 before.

IOI Performance Trajectories

- Strong correlation observed between consecutive IOI results.
- Significant variance in skill acquisition and unpredictability in single IOI results.
- Large spreads in improvement across consecutive IOIs.
- Reversion to the mean observed.
- Diminishing marginal improvements over multiple IOIs.

Geographic Analysis

- Population and HDI significantly correlated with aggregate IOI performance.
- Identification of regions performing better or worse than predicted by population and HDI.
- Competitive programming interest is a more critical factor.

Combining the Returning Contestant and the Geographical Analysis

- Explore interactions between contestant trajectories and country-level performance.
- Challenges due to limited data at this granularity.

Data Limitations (short sample period)

- Few datapoints complicate statistical tests and lead to large standard errors.
- Potential to revisit analysis with more data in future IOIs.

Incorporating Qualitative Analysis

- Combine statistics with qualitative insights for deeper understanding.
- Investigate IOI's role in global talent development and task approachability.