After, the actual performance may differ from the expected one. This means that the rating and volatility of a contestant did not predict the result correctly, and they shall be updated. Using (8) and (9), we may compute the rating—the *PerfAs*—that would correspond to this performance as shown in (10):

$$PerfAs_i = OldRating_i + CF * (AP_i - EP_i). \tag{10}$$

Weight of challenge for each player is calculated based on (11):

$$W_i = factor * \frac{1}{\left(1 - \left(\frac{0.42}{TimesPreviouslyPlayed_i} + 0.18\right)\right)} - 1, \tag{11}$$

where *factor* is 1.0 for contestants with ratings below 2000, 0.9 for contestants within ratings 2000-2500, and 0.8 for contestants with ratings above 2500.

Then, a cap—the maximum rating change for each player—is calculated using (12):

$$cap_i = 150 + \frac{1500}{TimesPreviouslyPlayed_i + 2}. \tag{12}$$

The new rating of each player is based on (13):

$$NewRating_i = \frac{OldRating_i + W_i * PerfAs_i}{1 + W_i}.^{3} \tag{13}$$

If the $cap_i < |NewRating_i - OldRating_i|$, then:

$$NewRating_i = OldRating_i \pm cap_i. \tag{14}$$

Finally, the new volatility of each player is calculated using (15):

$$NewVoltality_i = \sqrt{\frac{(NewRating_i - OldRating_i)^2}{W_i} + \frac{OldVoltality_i^2}{W_i + 1}}. \tag{15}$$

However, (15) is only used after the second game. The first new volatility is determined to be 385.

[3] $NewRating_i$ is the $PerfAs_i$ with the influence of weight of challenge W_i on the rating change.

4 Analysis

After applying the rating methods to the IOI datasets, section 4 evaluates the findings. Firstly, section 4.1 assesses the rating methods' performances, considering each's capability in predicting future contests' rankings in addition to proving the reliability of using the output as feedback for nations' skills. Secondly, section 4.2 shows how the results can be used to help nations enhance their skills and gather extra medals by studying Egypt as a case study.

4.1 Assessing the rating systems' performances

Table 2 unveils a sample showing the ability of each rating system to predict the rankings of the latest contest (2022) from the preceding overall ratings. Elo rankings and TrueSkill rankings are almost similar, and they are the closest to the actual rankings. Top Coder shows mostly correct predictions for the top 10 nations. Nevertheless, the system needs to be more precise to determine the exact rankings, particularly in the case of the Republic of Korea.

	Actual rankings of "Per Whole Contest" in 2022 contest	**Predicted rankings for 2022 contest by rating systems in "Per Whole Contest"**		
Rank	**Actual**	**Elo**	**TrueSkill**	**Top Coder**
1	China	China	China	China
2	USA	USA	USA	Russia
3	Japan	Russia	Russia	USA
4	Russia	Republic of Korea	Japan	Japan
5	Republic of Korea	Japan	Republic of Korea	Taiwan
6	Canada	Iran	Iran	Poland
7	Taiwan	Canada	Taiwan	Iran
8	Ukraine	Taiwan	Poland	Bulgaria
9	Iran	Singapore	Romania	Romania
10	Romania	Vietnam	Singapore	Republic of Korea

Table 2: A sample of the difference between the actual rankings of 2022 contest's "Per Whole Contest" category and the predicted rankings based on the previous contests' overall ratings by the three rating methods.

Table 3 shows the average predictive accuracy of each rating method for each category. From the table, we can declare which rating method is better for which category. The predictivities of Elo and TrueSkill are too close. Although Top Coder is mainly used for competitive programming, it was found to be the worst performing in the IOI case among the other rating systems. Compared to the results of Elo-MMR [9], our study shows impressive results according to pair inversion

predictivity in all the three rating systems. This proves the reliability of using the ratings as feedback for skill by nations' coaches to utilize them in improvement.

	Average Predictive Accuracy Per					
Rating system	Whole Contest	Graph Theory	Ad Hoc	Interactive	Data Structures	Others
Elo	**85.2189%**	78.5965%	**82.1554%**	74.6074%	**77.9353%**	**73.0332%**
TrueSkill	84.8023%	**78.6945%**	81.7188%	**74.8729%**	76.2757%	72.8895%
Top Coder	82.2446%	77.0277%	78.7000%	73.7930%	75.1334%	72.1569%

Table 3: The average predictive accuracy for each rating method in each category.

4.2 Case study: How can the ratings help Egypt improve?

The first significant question is which system's ratings to consider. The simple answer is all. We can rely on all of them as they all have high predictivities. Nonetheless, for each category, we prefer to consider the ratings of the system with the highest predictivity.

Figures 7, 8, and 9 illustrate the ratings of Egypt over the years in each category. From these figures, Elo maintains consistent volatility that is not too high or too low based on the scaling factor *k*. This advantages Elo because tuning the scaling factor *k* ensures that Elo runs at maximum efficiency. TrueSkill has high volatility during the first three rounds until it reaches consistent ratings. TrueSkill considers the first three rounds as the number required to reach stable ratings from non-stable ones. Once the stable condition is reached, the volatility gets too low. Top Coder shows high volatility during the whole period of ratings. This disadvantages Top Coder as the system sometimes gives too much unnecessary weight to some contests.

As mentioned in the introduction, the ratings will enable coaches to track their country's performance over the years, allowing them to identify improvement patterns. If we consider the Per Whole Contest Elo ratings because it has the highest predictivity, we will find that Egypt's rating significantly increased from 2019 to 2020. Hence, the coaches of the 2023 team can consult with the coaches from 2019-2020 to understand the actions taken to enhance the Egyptian team's performance. To sum up, helpful information can be inferred from the ratings by analyzing the patterns.

O date denotes the moment before the performance of ratings when all nations had initial ratings, Black (C) denotes Per Whole Contest, Orange (GT) denotes Per Graph Theory, Grey (AH) denotes Per Ad Hoc, Yellow (IN) denotes Per Interactive, Purple (DS) denotes Per Data Structures, and Green (OT) denotes Per Others

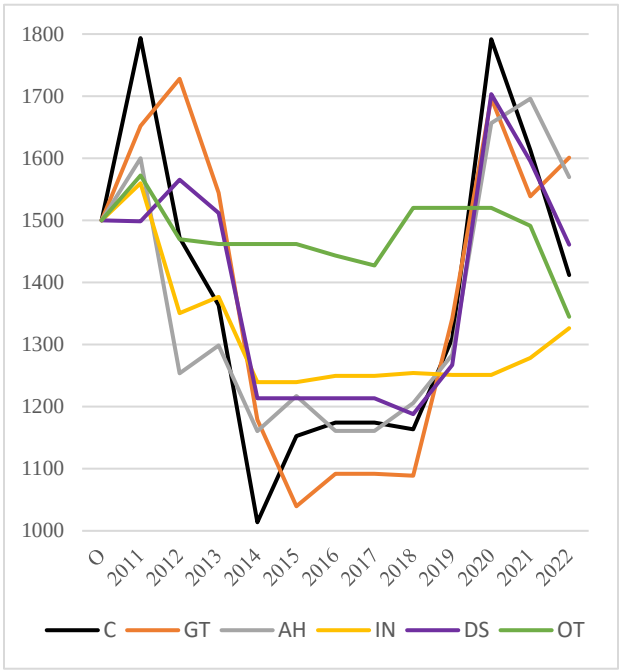


Figure 7: Egypt's Elo Ratings over years.

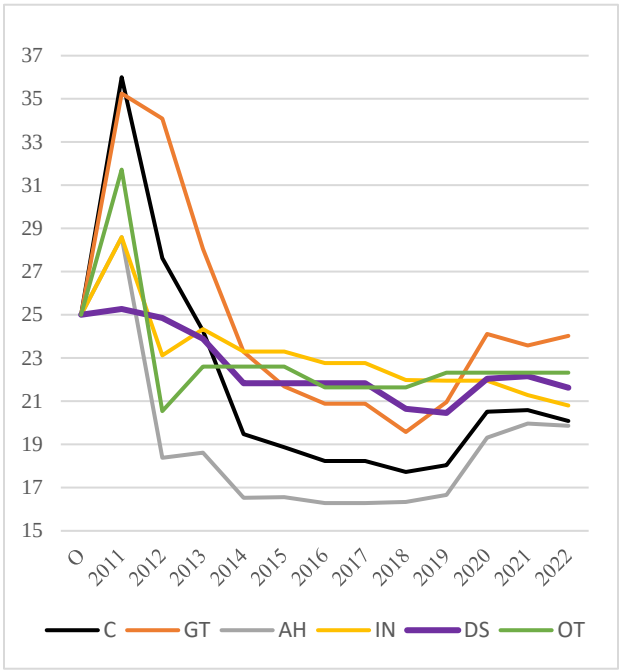


Figure 8: Egypt's TrueSkill Ratings over years.

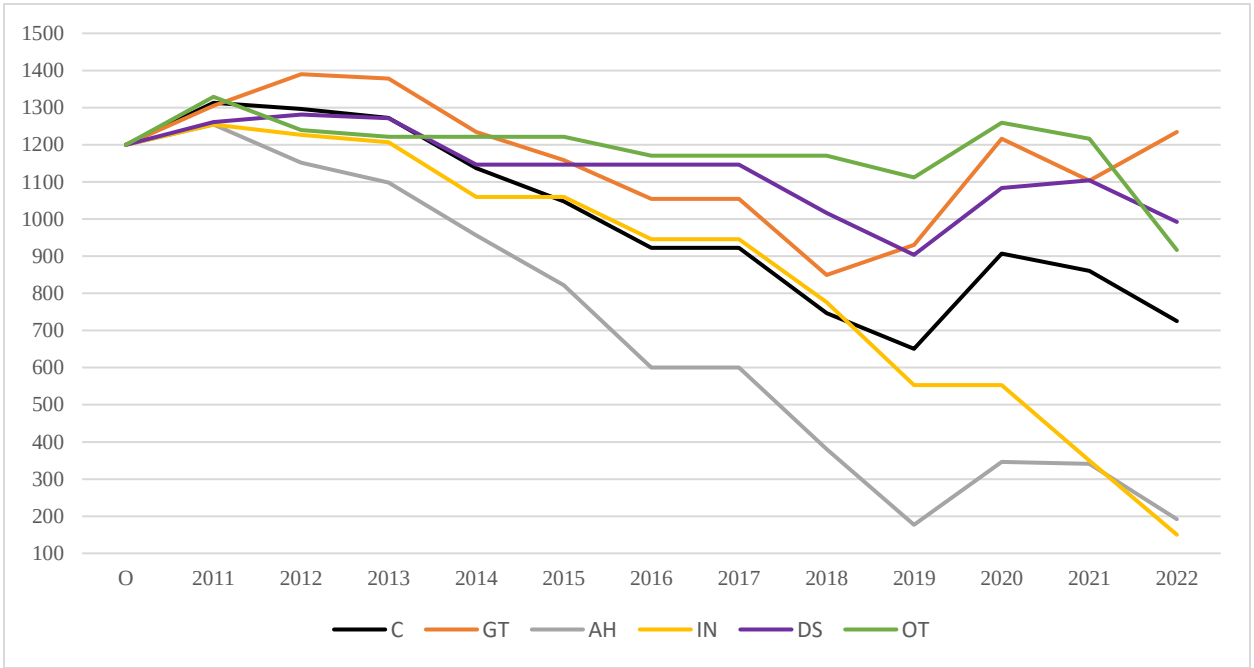


Figure 9: Egypt's Top Coder Ratings over years.

To assess the performance of Egypt or any other nation, we consider the rating system with the highest predictive accuracy for each category. Thus, comparing ratings as numbers will not reflect any helpful information as the scale of each rating system differs. For instance, without comparison charts, we cannot compare a score of 1500 on the SAT test and a score of 33 on the ACT test. In addition, the rating does not declare the performance of a nation, among others. For example, a nation with a rating of 2000, according to Elo in Per Graph Theory, might be in the top 10 percent among others, while a nation with a rating of 2500, according to Elo in Per Whole Contest, might be in the top 20 percent among others. Hence, standardization is crucial.

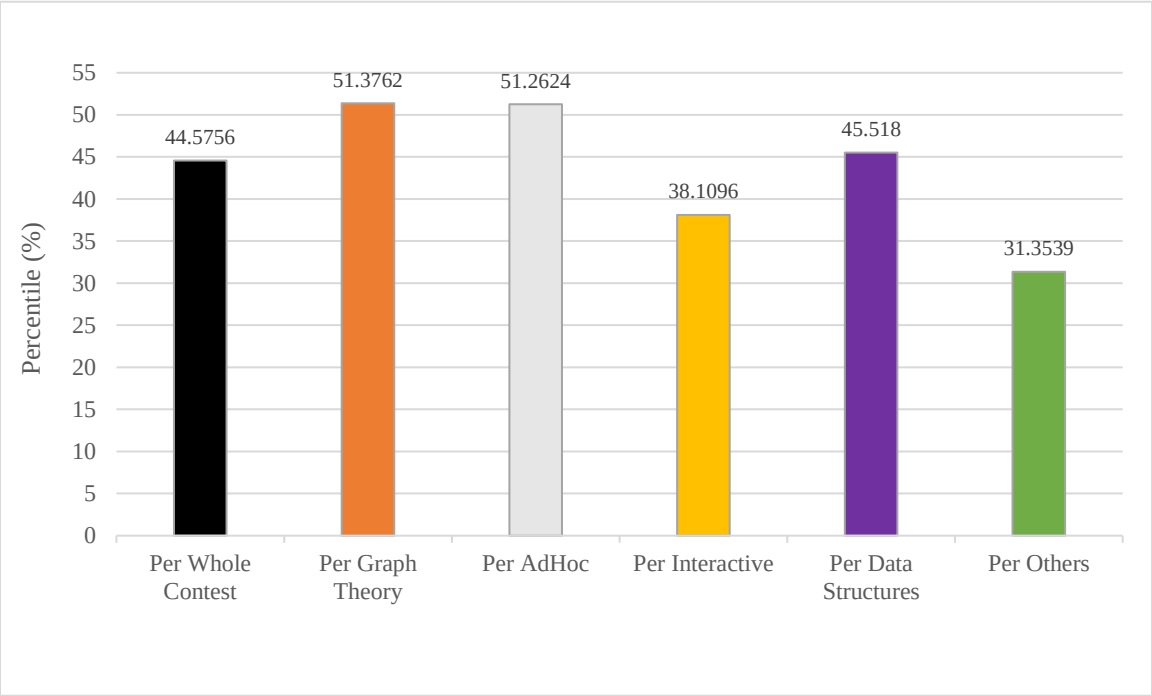
The standard score is the number of standard deviations in which a raw score is above or below the mean value of a sample of values. The standard score has many applications. One is standardizing scores of college tests such as the ACT and SAT. Since both have different scales, the *Z*-score helps in comparison by standardizing the scores. (16) represents the *Z*-score:

$$Z = \frac{x - \mu}{\sigma}, \tag{16}$$

Where Z is the standard score of a value of the data, x is the value of the data, μ is the mean, and σ is the standard deviation of the data. Standard scores help find the percentile, the value of which the data falls, of a sample, among others, based on the standard normal distribution. For example, if a team is at the 80% percentile, it is better than 80 percent of the teams. *Z*-scores can be turned into percentile using *Z*-score tables or calculation methods that rely on integral calculus. Using the Python statistics module to calculate the mean, standard deviation, and cumulative density function, we compare the output of the various rating methods by turning the *Z*-score into a percentile.

Figure 10 shows the percentile of Egypt in each category. Egypt is performing better at graph theory and ad hoc problems than other problems. As Egypt's percentile in whole contest rating is below 50 percent, Egypt is more unlikely to receive medals than other nations because, as mentioned earlier, only the top 50 percent receive medals. Nevertheless, by achieving consistent performance in the categories with the higher performances and concentrating on categories with lower performances, such as the interactive category, Egypt can reach a higher whole contest percentile, achieving more medals in the future.

Figure 10: The percentile of Egypt among other nations based on the latest ratings calculated after 2022 contest.



Nation	Percentile (%)	Rank	Gold	Silver	Bronze
China	98.25	1	4	0	0
USA	97.96	2	3	1	0
Russia	97.07	3	3	1	0
Japan	96.45	4	4	0	0
Republic of Korea	96.18	5	2	2	0
…					
India	77.26	21	0	2	2
Kazakhstan	72.95	22	0	2	2
Turkey	69.49	23	0	2	2
Brazil	67.66	24	0	3	1
Italy	65.84	25	0	1	3
…					
Switzerland	50.30	44	0	0	2
Macau	47.39	45	0	1	2
Latvia	46.24	46	0	0	2
Sweden	45.44	47	0	0	1
Egypt	**44.57**	**48**	**0**	**0**	**2**
Finland	44.32	49	0	0	2
Cuba	42.08	50	0	1	0
Mongolia	41.92	51	0	0	2
Cyprus	39.83	52	0	0	1
…					
Peru	29.49	61	0	0	1
Spain	29.41	62	0	0	1
Azerbaijan	27.84	63	0	0	1
Tajikistan	25.38	64	0	0	1
Norway	25.37	65	0	0	0

Table 4: A sample of medal achievements of few nations in 2022 contest. The ranks and percentiles are based on latest Per Whole Contest ratings after 2022 contest.

Finally, *Table 4* gives insights into the medal achievements of a few nations, showing their latest Per Whole Contest ranks and percentiles based on Elo in the last IOI 2022 contest. We can deduce from the table that nations within specific ranges of ranks nearly achieve the same number and types of medals. For instance, nations within ranks 44-52 achieve mostly two medals, and they are almost bronze. Hence, a nation needs to reach a lower range of ranks within which it can achieve more and better in type (gold > silver > bronze) medals. The ranges can be noticed from the full version of *Table 4* found on *"github.com/MoiMohamed/IOIRatings"*—the link comprises all the data, the codes of the rating methods, and the results.

5 Conclusion and Future Work

This paper proposed insights into improving and achieving more medals as a nation in the International Informatics Olympiad. By applying and comparing Elo, TrueSkill, and Top Coder rating methods to the IOI datasets between 2011 and 2022, we could analyze the performance of nations through ratings of whole contests and specify a rating for each problem category. The findings proved the reliability of rating methods in estimating nations' skills, as the rating methods showed high predictive accuracies. It was found that the performance of Elo and TrueSkill is nearly the same, with the highest predictive accuracies. Nevertheless, Top Coder, which is already being used in rating competitive programming contests, showed lower predictive accuracies.

By studying the patterns in rating changes over the years, the nations' coaches can utilize the ratings to know how to improve by assessing the periods of nations' improvement in performance. In addition, this study showed a method to compare the performances of each nation in each problem category based on the concepts of *Z*-scores, percentiles, and standardizing. A nation can assess its performance by observing its percentile in each category among other nations. Thus, a nation can work to improve in the categories that weaken their final Per Whole Contest ratings. From relating percentiles to medal achievements, we deduced that each specific range of percentiles achieved nearly the same number and types of medals. Hence, these ranges can be used to know how much improvement in percentiles is needed to achieve more and better medals.

Finally, we recommend a sensitive analysis of the results for future work to find which categories are easier to improve. Moreover, we recommend running the

same experiments using more advanced versions of Elo, such as Elo-MMR, as Elo's modifications can help achieve better predictive accuracies.

6 Bibliography

[1] IOI Statistics, https://stats.ioinformatics.org/.

[2] T. VERHOEFF, 20 years of IOI competition tasks - ioinformatics.org, https://ioinformatics.org/journal/INFOL047.pdf.

[3] "IOI '22 P1 - Catfish Farm - DMOJ: Modern online judge," DMOJ, https://dmoj.ca/problem/ioi22p1.

[4] R. D. and N. P. Justin Moore, "Who's the best formula One driver of all time?," FiveThirtyEight, https://fivethirtyeight.com/features/formula-one-racing/.

[5] R. Herbrich, T. Minka, and T. Graepel, "TrueSkillTM: A bayesian skill rating system," *Advances in Neural Information Processing Systems 19*, pp. 569–576, 2007. doi:10.7551/mitpress/7503.003.0076

[6] M. Vojnovic, *Contest Theory - Incentive Mechanisms and Ranking Methods*. Cambridge University Press, 2016.

[7] M. E. Glickman and A. C. Jones, "Rating the chess rating system," *CHANCE-BERLIN THEN NEW YORK*, vol. 12, pp. 21–28, 1999.

[8] R. Lehmann and K. Wohlrabe, "Who is the 'Journal Grand Master'? A new ranking based on the Elo rating system," *Journal of Informetrics*, vol. 11, no. 3, pp. 800–809, 2017. doi:10.1016/j.joi.2017.05.004

[9] A. Ebtekar and P. Liu, "Elo-MMR: A rating system for massive multiplayer competitions," *Proceedings of the Web Conference 2021*, 2021. doi:10.1145/3442381.3450091

[10] Top Coder Rating System, https://www.topcoder.com/thrive/articles/Ratings.

[11] M. Forišek, *Theoretical and Practical Aspects of Programming Contest Ratings*. [Online]. Available: https://people.ksp.sk/~misof/publications/2009thesis.pd