Thanks for liking the structure, motivation and technique of our paper :)

About your question about the intention of applying reinforcement learning instead of supervised learning techniques, we mentioned in both introduction and the cover letter that player ranking is essentially an unsupervised problem, as its ground truth is not available.

That's the reason why we apply a semi-supervised method reinforcement learning.

We assume by saying 'use a fully supervised technique to model ...', you mean using the goal indicator (reward of RL, discussed in Play Dynamic in NHL section) as supervision value. But in that case, we are predicting the scoring chance, instead of evaluating the influence of player's actions to score the next goal, which is naturally achieved by Q function, as it will look ahead to the next goal. And as you have talked about interpretability, yes, Q value does provided insight into the influence of player's action under different game situations by looking ahead to the next scoring event.

About your question of 'why is a deep network required in this application', we do experiment about applying a shallow network containing single hidden layers. But our result shows that it doesn't converge very well (completely diverge sometimes). It's famous issue raised by the paper 'Analysis of temporal-difference learning with function approximation'. But Peter Stone managed to overcome it with deep recurrent network in his recent work 'Deep recurrent q-learning for partially observable mdps'. That's why we follow his structure and further extend it to on-policy setting and add dynamic trace length, as it's discussed in section 5. Our experiment proves the effectiveness of this Deep Reinforcement Learning structure.

For your question about 'how this approach would generalize beyond the games that have small number of players and player takes the "impact" actions ', we mentioned in section 4 that our data set contains 2,233 NHL player. It's a large number, as NHL is the largest Ice Hockey league in North America. And as our Q function can evaluate all the actions (see our paper and the fourth reviewer's comment), "impact actions” indicates all the actions in dataset.

Thanks for your encouraging comment. We did adsorb some cool ideas from baseball. Their Goal-Above-Replacement (GAR) and Win-Above-Replacement have been applied as comparison methods in our paper. And we will put our source code on github and let others to try our methods. For the typos, we will fix them in our final version of paper.

Thank you for encouraging comment. We have extended the application of deep reinforcement learning to player evaluation and designed empirical evaluation to prove the effectiveness of our GIM metric.

Thank you for the encouragement and for engaging with many interesting and important details. This is one of the most helpful reviews we have received!

For our following work [Hausknecht and Stone, 2015], we mentioned in the introduction and related work that we extend it to on-policy setting and add the dynamic trace length.

For your question about the terminology of reward, as you said, yes, Q\_team is a generic definition for Q\_home, Q\_away and Q\_neither, we have discussed it in the section 4. And weights for all Q functions are learned together, as it's shown in Fig.3 (there are three output nodes). We will add our description about them in our final version of paper.

About your question about ' why the \*change\* in Q value is used', we talked about it in section 6.1. It's because we want to measure the quality of an action by how much it changes the expected return of a player’s team. It's the motivation of using difference between Q values. And we have try to use a Q value to directly evaluate the players' action, but result doesn't look well in the Empirical Evaluation section, that's another intention to invent our new GIM metric.

For your concern about ' adding the discussion about why the low correlation (of all metrics) for some statistics', it's a good point and we will add it in the final version of our paper. The primary explain is because the number of event that related to some of the standard success measures is rather small in our dataset. Further experiments are required to prove this idea.

And about your question about 'whether the teammates have a role in individual performance ', the answer is yes and it is the motivation of using recurrent neural networks LSTM as it's descried in the section 4. To produce a Q values for a player, LSTM will trace a series of events, which contain actions of his teammates. We will clarify it more in revising the paper.