# Rebuttals (2000 characters)

## Reviewer 1

Thank you for liking the structure, motivation and technique of our paper. About the background, the reviews have given us good guidance for what to add and what to cut down.

“Why not use a supervised technique to model the relationship between actions, observations and rewards?” Value function learning in RL is an unsupervised problem because the Q-function models the relationship between actions, observations and \****expected future reward\****. This is clear when modelling the probability of winning at the end of the game the model looks ahead not to the next event, but to an expected value over all possible future trajectories. Our Q-function similarly looks ahead to predict which team will score the next goal, within an unbounded number of steps. It does \****not\**** predict the goal indicator g\_t (= reward).

“Why is a deep network required?” We experimented with a network. Over many settings, typically weight training diverges, or converges slowly to a Q-function with poor data fit. NN divergence is a famous issue (http://www.mit.edu/~jnt/Papers/J063-97-bvr-td.pdf 1997). The recent deep recurrent network in our reference [Hausknecht and Stone] overcomes divergence. We built on their structure and extended it as discussed in Sec. 5 (on-policy setting, dynamic trace length). Our evaluation shows that the Q-function learned in this architecture is effective for player ranking.

“This seems well-suited when there are small number of players, and one player is likely to take the impact action.” We rank 2,233 NHL players (see Sec. 4). If you mean that at any game time, only a small number of players are on the ice, that is true of most sports. In a dataset that records only few actions assigned for each player (e.g. only goals), one can use regression to assess how important the presence of a player is compared to others (e.g. Schuckers ref). Combining regression with our goal impact metric is a promising direction for future work that would generalize our approach to more datasets.

## Reviewer 2

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

## Reviewer 3

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

## Reviewer 4

Thank you for the encouragement and for engaging with the details. One of the most helpful reviews we have received!

About [Hausknecht and Stone, 2015]: Similarities include: a) same NN architecture b) using LSTM to address partial observability. Differences include: a) dynamic trace length based on possession (Sec.5), b) on-policy setting (Intro) rather than off-policy control.

About the rewards terminology, yes, they are sparse, and Q\_team is a generic definition. We will revise this terminology. And weights for all Q functions are learned together, Fig.3 shows three output nodes.

We did experiments with Q-values directly as action values. This metric shows weaker correlations with standard success metrics. Intuitively, this is because there are two reasons why Q(s,a) can be high: i) s can be a high-value state (e.g. in front of empty net), or ii) the action increases the scoring chance above the s value. In case i), the player deserves no special credit. Delta-Q cancels the state effect to focus on ii). Our predecessor papers also use a delta approach [Cervone and Schulte refs].

We will follow your suggestion to “add the discussion about why the low correlation for some statistics”. The primary explanation is that those success measures have only a weak relationship with scoring, and occur rarely. Further investigation is required to quantify this phenomenon.

Our results do suggest that the available information is sufficient to evaluate action impact in terms of the probability of scoring the next goal. While complete tracking data about teammates would improve the Q-predictions, this lack is partially compensated by using the event history via an LSTM. Hausknecht and Stone also found that the LSTM can compensate for partial observability.

Yes TGP should be TOI.

The neural net learns Q-values on the **whole** dataset (e.g. action sequences near opponent’s goal are assigned high scoring chances.) Fig.4 applies this general NN to the histories in a specific game.