Thank you for all your comments. Our responses are detailed below, and we will incorporate them in the final paper.

Reviewer 1: Variable number of comparisons k. To simplify notation and analysis, it's standard in the BTL literature to assume that pairs are compared a fixed number k times (see [10] or [12]). As explained in lines 150-152, "if k varies between players so that i and j play $k_{i,j} = k_{j,i}$ games," the data Z can be normalized accordingly for our algorithm. The bandwidth h can be altered, e.g., to use $k' = \min_{\{i,j\}\in\mathcal{G}(n,p)} k_{i,j}$ in place of k in equation (8). This would yield corresponding theoretical guarantees with k'. We will clarify this in the final paper. One could also carry out the long analysis in [10] (used in Lemma 3) and our analysis while keeping track of the $k_{i,j}$'s, and derive a corresponding h that yields finer bounds. We decided to omit this due to space constraints.

Reviewer 1: Ad-hoc steps in experiments. We have already explained all the details needed to run all our experiments in the paper. Perhaps the reviewer missed these details. Please see lines 150-151 and 238-246 for the precise values and choices used to execute our algorithm, e.g., $h = 0.3n^{-1/4}$. On the other hand, we had not mentioned that we chose the constant 0.3 in h, and the level of smoothing ("…game is counted as 20 games …"), by eyeballing when the densities in Figure 1 looked 'smooth.' Moreover, our qualitative results and trends in section 4 remain the same for a range of values around 0.3 and 20 (e.g., 0.4 or 30). We will clarify these points in the final paper.

Reviewer 1: Future directions and minor comments. The minor corrections and clarifications will be made in the final paper. Thank you very much for suggesting future research avenues pertaining to Bayesian or MDL approaches and density estimation based on mutual fund performances.

Reviewer 2: Interpreting skill PDFs. We assume a lower bound on our skill PDFs over $[\delta, 1]$ in a neighborhood of 1 (see line 114). This implies that $\max_i \alpha_i \approx 1$ with high probability for large n (see (31) in supplementary materials). Intuitively, we are re-normalizing all skill parameters so that the maximum one is essentially 1. So, if there are just two teams with skills δ and 2δ, these values will be re-normalized to 0.5 and 1. Since $\max_i \alpha_i \approx 1$, it is reasonable for the uniform skill PDF to put more mass on larger intervals, i.e., the uniform skill PDF is interpretable. Thus, we do not see any immediate advantage of using logits. In Figures 1c and 1f, "World" and "English" have different skill PDFs but similar skill scores, because different PDFs can have the same KL divergence to the uniform PDF. This artifact remains even if we use logits. On a separate note, since the logits $\omega_i = \log(\alpha_i)$ are i.i.d. with PDF $P_\omega(t) = e^t P_\alpha(e^t)$, when δ is constant and $h < \delta$, we have an estimator $\widehat{\mathcal{P}}_2^*(t) = e^t \widehat{\mathcal{P}}^*(e^t)$ for P_ω if desired. By substitution, $\mathbb{E}\big[\int(\widehat{\mathcal{P}}_2^* - P_\omega)^2\big]$ $\leq (1+\delta)\,\mathbb{E}\big[\int(\widehat{\mathcal{P}}^* - P_\alpha)^2\big]$. Therefore, the upper bound in Theorem 3 also holds for MSE estimation of logit PDFs.

Reviewers 3 & 4: Why not estimate skill score directly from $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ (instead of estimating skill PDF P_α)? We outline several reasons to estimate P_α: (i) If one seeks to estimate a specific functional of P_α, e.g., moments or variance, it is possible to estimate this directly from $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$. (This is still nontrivial because careful analysis is needed to prove consistency of estimation based on noisy pairwise comparisons.) However, P_α *contains information about all such functionals* and provides a lot more qualitative information as shown in Figure 1. Since different functionals are needed for different applications, a good estimate of P_α rather than just samples is very useful. Our main contribution is showing that the entire smooth density P_α can be estimated from $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ as well as if we had access to the true $\alpha_1, \ldots, \alpha_n$. (ii) The dual characterization of TV distance and the Cauchy-Schwarz inequality give: $T \triangleq \sup_{P_\alpha, \|f\|_\infty \leq 1} \mathbb{E}\big[(\int f \mathrm{d}\widehat{\mathcal{P}}^* - \int f \mathrm{d}P_\alpha)^2\big] \leq \sup_{P_\alpha} \mathbb{E}\big[(\int|\widehat{\mathcal{P}}^* - P_\alpha|)^2\big] \leq 3 \sup_{P_\alpha} \mathbb{E}\big[\int(\widehat{\mathcal{P}}^* - P_\alpha)^2\big]$, where the first sup is over P_α and all functions f bounded by 1. Thus, the bound in Theorem 3 holds for T. So, by estimating P_α, we obtain *uniform guarantees on estimating any bounded statistic* of the form $\mathbb{E}[f(\alpha)]$, which includes all moments. (iii) We believe differential entropy $h(P_\alpha)$ is an excellent overall skill score, and standard non-parametric estimators for it in the literature (e.g., integral, resubstitution, or splitting data estimators) *require an estimate of* P_α first to plug in. (Also, quantization theory shows that discrete "entropy of the empirical distribution" of $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$, with uniform binning, is a poorer estimate of $h(P_\alpha) - \log(\text{bin size})$.) (iv) Philosophically, we believe that skill levels of players, like height or weight, exhibit a distribution of values, and a tournament contains samples from this distribution. Hence, the "right" skill score is based on P_α rather than the realizations $\alpha_1, \ldots, \alpha_n$. We will elaborate on these points in the final paper.

Reviewer 3: "Skill" vs. "exciting" vs. "competitive". We will clarify that the differential entropy based "overall skill score" measures variation of skills, not the actual skills of players, in the final paper. We used "exciting" sparingly in the paper, but agree that it may not precisely capture what we mean. So, we will change or clarify it in technical discussions in the final paper. We have also sparingly referred to tournaments with high overall skill scores as "competitive," because many teams have similar skill parameters and game outcomes are less predictable. This usage seems reasonable and we have retained it. We also agree that closely matched teams may have many non-competitive games, and our overall skill score is indeed an "average measure" which may not capture these low probability events.

Reviewer 4: Theorem statements. Using the phrase "sufficiently large" is *standard practice* in mathematical statistics (see, e.g., [10]), and it has a very *precise meaning*. "Sufficiently large n (or constant c)" means that "there exists a constant A such that for all $n \geq A$ (or $c \geq A$)." Here, the values of A may depend on other constant problem parameters, and they can be deduced from our proofs. For example, in Theorem 1, the constant $c_{15} \geq 2c_4/(\delta\sqrt{pk})$, where c_4 is the universal constant from Lemma 3 (which is Theorem 3.1 in [10]). Moreover, for "large constants," we already mention which parameters A depends on in the paper. Since we do not derive sharp values of A, it is not illustrative to include them in theorem statements. However, our theorem statements are rigorous; e.g., they directly imply big-O style results.