A DATASETS

We evaluate the performance of ShaRP and compare it to other local feature importance methods, using several real and synthetic datasets, with the corresponding ranking tasks. Dataset properties, along with ranker type (score-based or learned) are summarized in Table 1 and described below. We show the relationship between score and rank for score-based ranking tasks in Figure 9.

CS Rankings (CSR) ranks 189 Computer Science departments in the US based on a normalized publication count of the faculty across 4 research areas: AI, Systems (Sys), Theory (Th), and Interdisciplinary (Int) [35]. We use publication data for 2013-2023, with the scoring function provided by CSRankings, a geometric mean of the adjusted counts per area, with # of sub-areas as exponent:

$$f = \sqrt[27]{(AC_{AI}^{5}+1)(AC_{Sys}^{12}+1)(AC_{Th}^{3}+1)(AC_{Int}^{7}+1)}$$

ATP Tennis (ATP) is based on publicly available 2020-2023 performance data of tennis players from the Association of Tennis Professionals (ATP) [19]. We use 2022 data that includes 5 performance-related attributes of 86 players. We select 2022 because this is the year in which data for all 5 attributes is available for the highest number of players. We use the following scoring function that we recovered from the ATP site using the scores:

$$f = 100 \times (\%\ \text{1st Serve}) + 100 \times (\%\text{1st Serve Points Won}) + 100 \times (\%\text{2nd Serve Points Won}) + 100 \times (\%\ \text{Service Points Won}) + 100 \times (\text{Avg Aces/Match}) - 100 \times (\text{Avg Double Faults/Match})$$

Times Higher Education (THE) is a dataset of worldwide university rankings [18]. It contains the university name, country, and the scores assigned to the university by Times Higher Education for teaching (TEA), research (RES), citations (CIT), income (INC), and international students (INT). We use 2020 data, for consistency with Anahideh and Mohabbati-Kalejahi [2] who also used it in their paper, with the scoring function provided by THE:

$$f = 0.3 \times TEA + 0.3 \times RES + 0.3 \times CIT + 0.025 \times INC + 0.075 \times INT$$

Moving Company. The moving company scenario [37] simulates a hiring process where job applicants are ranked based on their *qualification score*, computed as a function of their weight lifting ability, sex and race. We train two different rankers, over two scenarios:

(1) Using the original data from a previous hiring process from that company, where female applicants generally display lower weight-lifting ability than male applicants and a lower qualification score. In addition, black applicants have a lower qualification score compared to white applicants, but similar weight-lifting ability. Hence black females face greater discrimination compared to the rest of the applicants.
(2) After applying the intersectional fairness intervention proposed in the same paper over the data.

All versions of this dataset (both scenarios and train/test sets) contain 2000 tuples.

We use an eXtreme Gradient Boosting (XGB) and a Light Gradient-Boosting (LGB) Machine to model the rankings of the applicants in the training set, and infer and compute the feature contributions of the applicants in the test set, i.e., all results reported in this section correspond to the test set. The XGB ranker was defined with the pairwise ranking objective, while the LGB ranker uses the lambdarank objective.

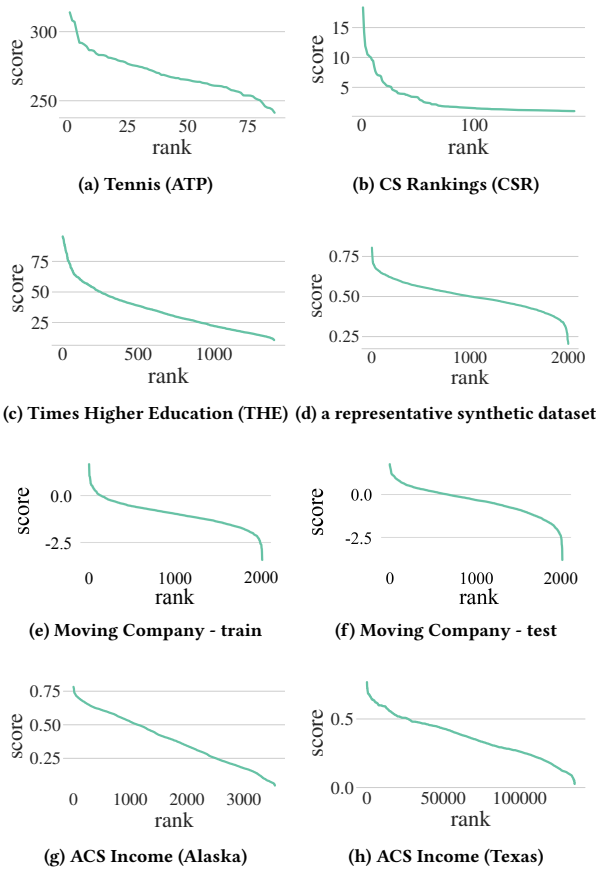


Figure 9: The relationship between an item's score (y-axis) and its rank (x-axis) for four score-based tasks.

Synthetic datasets. We also use numerous synthetic datasets to showcase specific quantitative and qualitative aspects of local feature-based explanations and metrics, and to study specific aspects of performance. These datasets contain 2,000 tuples. In five of them, items have 2 features, x_1 and x_2, distributed according to the uniform, Gaussian, or Bernoulli distributions, with varying parameters. We experiment with both independent and correlated features. Each synthetic dataset consists of 2, 000 items. We use three linear scoring functions: $f_1 = 0.8 \times x_1 + 0.2 \times x_2$, $f_2 = 0.5 \times x_1 + 0.5 \times x_2$, and $f_3 = 0.2 \times x_1 + 0.8 \times x_2$.

To explore correlations further, we create three datasets that have three Normal features x_1, x_2, and x_3, and 2,000 items. In the first dataset, all features are independent. In the second, we draw x_1 and x_2 from the 2D Gaussian and they are negatively correlated with correlation -0.8. The third feature x_3 is independent. For the third dataset, we draw the features from the 3D Gaussian. x_1 and

x_2 are negatively correlated with correlation -0.8, x_1 and x_3 are positively correlated with correlation 0.6, and x_2 and x_3 are negatively correlated with correlation -0.2. For all three datasets, we use the same scoring function $f_4 = 0.33 \times x_1 + 0.33 \times x_2 + 0.34 \times x_3$.

B DISTRIBUTIONAL ANALYSIS FOR RANKING

Fixed scoring function, varying data distribution. In this experiment, we illustrate that feature importance is impacted by the data distribution of the scoring features to a much greater extent than by the feature weights in the scoring function. Further, we show that feature importance varies by rank stratum. In Figure 10, we show rank QoI for 4 synthetic datasets with the same scoring function f_2.

We observe that, while the features have equal scoring function weights, their contributions to rank QoI differ for most datasets. In D_1, the Bernoulli-distributed x_2 determines whether the item is in the top or the bottom half of the ranking, while the Gaussian-distributed x_1 is responsible for the ranking inside each half. For D_2, the uniform x_1 has higher importance because it often takes on larger values than the Gaussian x_2. In D_4, x_1 and x_2 are negatively correlated, so when one contributes positively, the other contributes negatively. Only for D_3, with two uniform identically distributed features, the median contributions of both features are approximately the same within each stratum.

Additionally, we see that feature contributions differ per rank stratum. For example, for D_3, the medians show a downward trajectory across strata. This is because they quantify the expected change (positive or negative) in the number of rank positions to which the current feature values contribute. Also for D_3, feature contributions have higher variance in the middle of the range, because a 40-60% rank corresponds to many feature value combinations.

Fixed data distribution, varying scoring function. In this experiment, we investigate the impact of the scoring function on rank and top-k QoI for two datasets. In Figure 11, we use D_3 and see that the contributions to rank QoI vary depending on the scoring function. For f_1, x_1 is the only important feature (although it carries 0.8 — and not 1.0 — of the weight). This can be explained by the compounding effect of the higher scoring function weight and higher variance of the distribution from which x_1 is drawn. Between f_2 and f_3, features x_1 and x_2 switch positions in terms of importance, and show a similar trend, despite being associated with different scoring function weights (0.5 & 0.5 vs. 0.2 & 0.8). This, again, can be explained by the higher variance of x_1, hence, x_2 needs a higher scoring function weight to compensate for lower variance and achieve similar importance.

Top-k access. Access to the top-k is determined by the interaction between the scoring feature weights and the distributions of these features. The top-k QoI tells us how important each feature is when we consider only access to the top-k. A positive feature contribution signifies that changing the feature's value will result in decreased chances of getting to the top-k. A very high (or very low) value shows that the changes are significant. Figure 12 illustrates this for datasets D_2 and D_3. When we consider two identical uniform features that have equal weights (D_2 under f_2), we first notice that their control of top-k access is identical as expected. Additionally, we see that for the top-10, changing either feature would reduce access to the top-k (the values are both very positive). However, for each stratum up to the top-70%, changing either feature can contribute either positively or negatively.

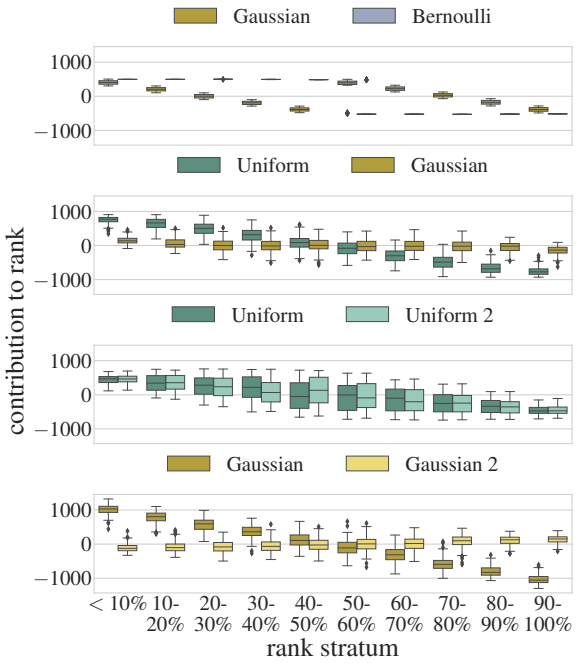


Figure 10: The rank QoI using $f_2 = 0.5 \times x_1 + 0.5 \times x_2$ for four datasets; D_1: $x_1 \sim N(0.5, 0.1)$, $x_2 \sim Bern(0.5)$; D_2: $x_1 \sim [0, 1]$, $x_2 \sim N(0.5, 0.1)$; D_3: $x_1 \sim [0, 1]$, $x_2 \sim [0, 1]$; D_4: $x_1 \sim N(0.5, 0.05)$, $x_2 \sim N(0.75, 0.016)$, with -0.8 correlation. Feature contributions are different per rank stratum and data distribution.

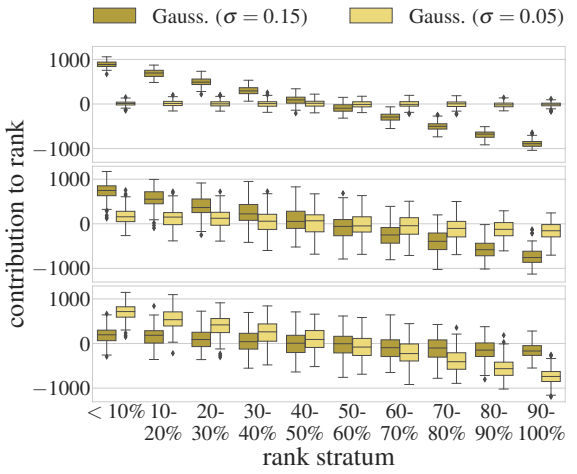


Figure 11: Rank QoI for D_5: $x_1 \sim N(0.5, 0.1)$, $x_2 \sim N(0.5, 0.05)$. Subplots correspond to different scoring functions: $f_1 = 0.8 \times x_1 + 0.2 \times x_2$ (top), $f_2 = 0.5 \times x_1 + 0.5 \times x_2$ (middle), $f_3 = 0.2 \times x_1 + 0.8 \times x_2$ (bottom).

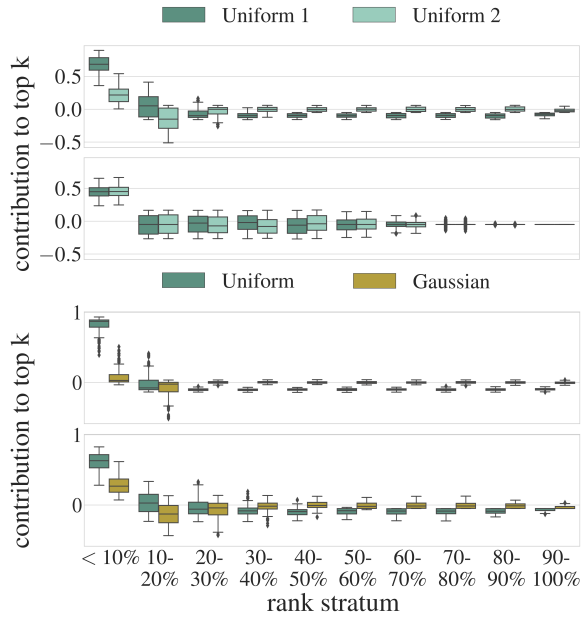


Figure 12: Top-k QoI for $k = 10\%$, D_2: $x_1 \sim [0,1]$, $x_2 \sim N(0.5, 0.1)$, and D_3: $x_1 \sim [0,1]$, $x_2 \sim [0,1]$. Subplots correspond to different scoring functions: $f_1 = 0.8{\times}x_1{+}0.2{\times}x_2$ (top), $f_2 = 0.5{\times}x_1{+}0.5{\times}x_2$ (bottom).

When we consider two identical uniform features (D_1) one of which has a higher weight (f_1) or dataset D_2 (under either f_1 or f_2), we see again that for the top-10, changing either feature would reduce access to the top-k. Also, we see that how high the contributions are depends on the distributions. However, we see that for the top 10%-20% changing the second, less important feature would increase the chances of getting into the top-k. For the rest of the strata, with some variations depending on the dataset and function, changing the most important feature provides non-zero probability of moving to the top-k and interestingly, this persists even for the lower strata. Evidence that items from lower strata can move to the top-k under some scoring functions and feature distributions counters the assumption of Anahideh and Mohabbati-Kalejahi [2] that changes in rank are localized.

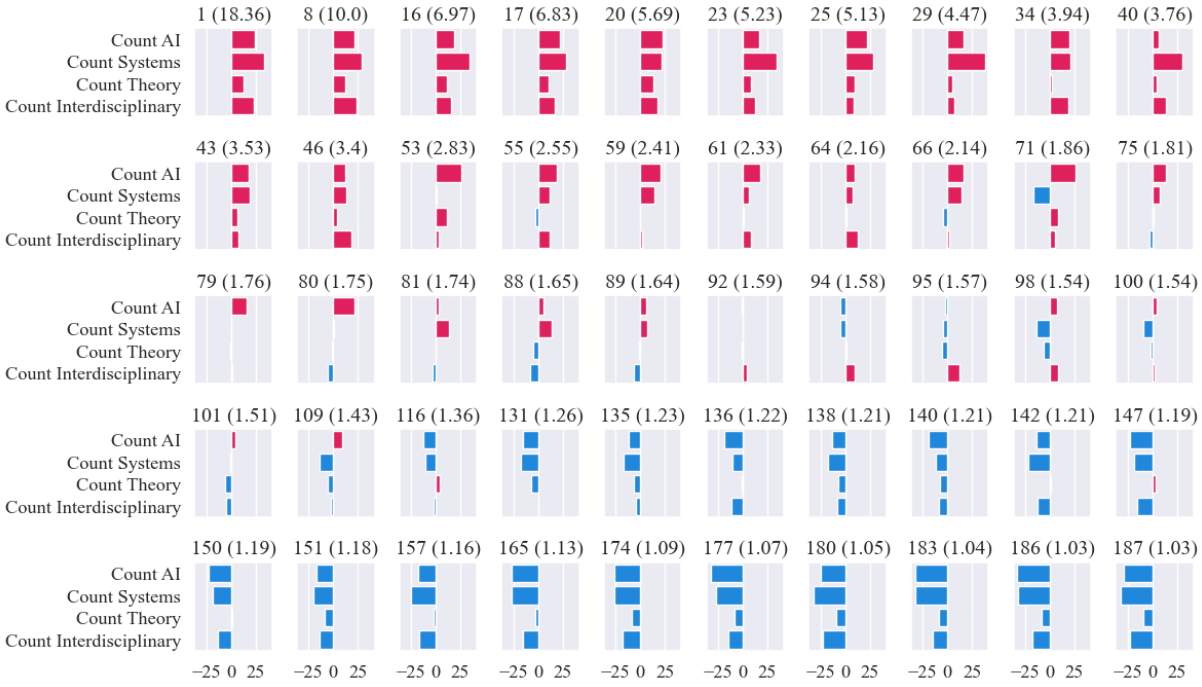
C RANK-QOI-BASED AND SCORE-QOI-BASED EXPLANATIONS FOR CS RANKINGS

In Section 4 we discuss the differences between the rank QoI and the score QoI for the CSRankings dataset. In this section we provide additional details for this comparison. In Figure 13 we provide local Shapley value explanations for fifty universities from the CSR dataset for both the rank (Fig. 13a) and the score QoI (Fig. 13b). These universities are randomly chosen, they are approximately 25% of the dataset and span the entire ranking. Each subplot in each subfigure shows one explanation for one university, its title shows its rank and its score (the score is in parentheses). The universities are the same across both subfigures. Looking at this collection of explanations, we can see how the rank and the score QoI behave significantly differently. Matching what we showed in Figure 3, the score QoI explanations become indistinguishable as we move down the ranking. Additionally, the contributions of all features becomes negative around rank 61 for the score QoI as opposed to 131 for the rank QoI. Finally, for the score QoI the contributions are very small for almost all universities as opposed to the rank QoI where the contributions are small for the middle of the ranks. There are two main reasons why the behavior between the rank and the score QoI based explanations is so different. The first is that the score-to-rank relationship is exponentially decreasing for this dataset (see Fig.9b). The second is that Shapley values explain the contribution of each feature to the distance of the outcome from the mean outcome. The mean score for this dataset is 2.72 and its range is 18.36-1.03, while the mean rank is 95 and its range 1-198. Together, these two facts mean that for the score QoI, for most items, the distance between its score and the mean score is very small. This is why the contributions are very low for most items. Additionally, the mean outcome is very influenced by the outliers at the top and most items have negative contributions for all their features even when ranked at the top 30% (e.g. university ranked at position 61). As discussed in multiple works that compare explainability methods and explore their desired properties, for instance [4, 24] explanations should differ when the outcomes and the items are different. In these figures, we can see that this is not the case for the score QoI based explanations. This behavior of the local explanations, coupled with the fact that the score QoI is not able to capture when the rank changes (see Section 1), argue for using the rank QoI when explaining rankings.

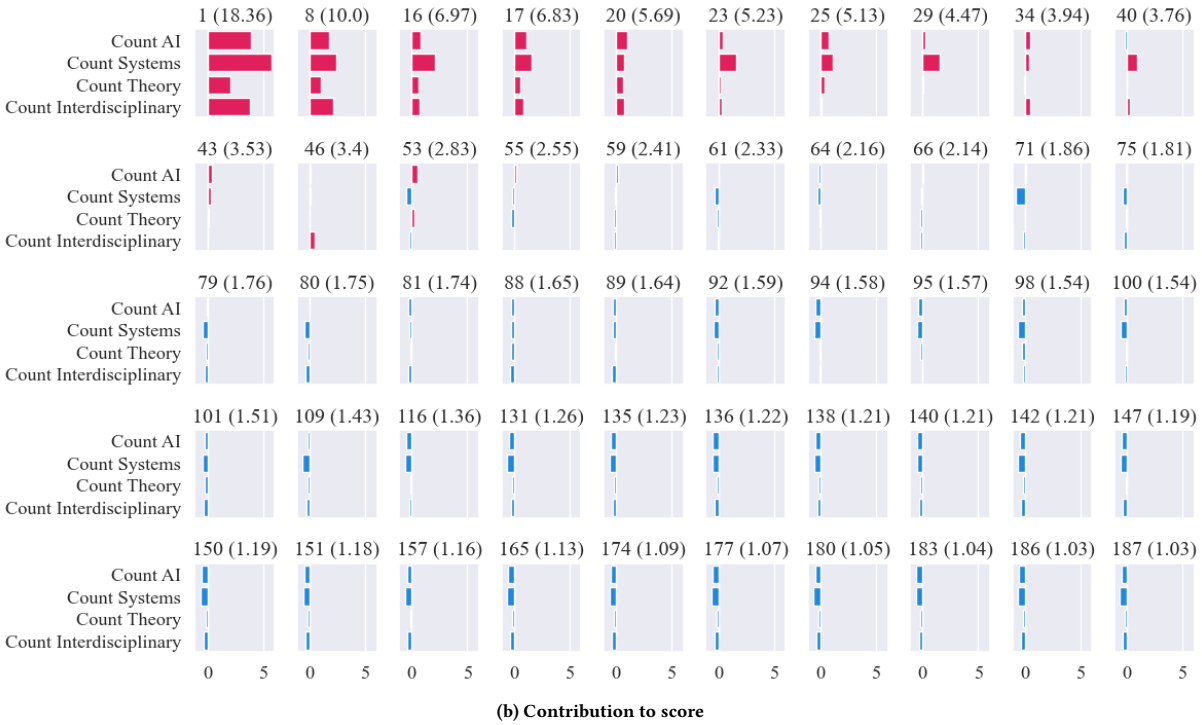
D IMPLEMENTATION OF HIL

HIL [39] is the only other method that recommends the usage of ranks as a profit function for individual explanations in ranking. While this method is not general, we are interested in comparing it with our rank QoI. This was not straightforward because the method is available as a web app that works only for linear weight scoring functions and datasets of two Gaussian features. To compare the rank-relevance contributions introduced in that paper to the rank QoI, we adapted their method using their definitions and code. This implementation is available alongside our own. Further, we extended their method to work with the specific non-linear scoring function used by CS Ranking, by changing the way that Std rank and Std score (discussed below) are computed.

More specifically, because HIL [39] works only with linear weight scoring functions, they do not provide a full Shapley values implementation but use the linear weights to approximate Shapley values assuming feature independence, see Corollary 1 in [21] and also [34]. This is a well-established method to compute Shapley values for linear weights also implemented by SHAP, so we do not compare with this part of the method. In addition, HIL defines two methods to acquire feature contributions: "standardized Shapley values" and "rank relevance Shapley values," which we will call Std score and Std rank, respectively. Those are not calculated using the linear weight method described above, but rather directly from the weights, and without using the mean score or rank. For an item $\mathbf{v}$, each feature i contribution for Std score is $\phi_i = \frac{\beta_i \mathbf{v}_i}{\sum_{\mathbf{u}\in\mathcal{D}} f(u)}$, where β_i is the weight for feature i. In other words, the contribution of each feature for each item is the score contribution of this feature over



(a) Contribution to rank



(b) Contribution to score

Figure 13: Shapley value explanations for fifty random universities for the rank QoI and the score QoI. The title of each subplot indicates the rank of each university and also contains its score in parentheses. The exponentially decreasing score-to-rank relationship and the dependence of Shapley values on the mean make score explanations indistinguishable and negative for most of the items.

the sum of all scores for all items. Similarly, for Std rank, the contribution of feature i for an item $\mathbf{v}$ is $\phi_i = \beta_i \mathbf{v}_i \alpha_{\mathbf{v}}$, where $\alpha_{\mathbf{v}}$ is a scaling factor used to transform the score of the specific item to the rank of the specific item calculated as $\alpha_{\mathbf{v}} = \frac{(\max_{r \in r_{\mathcal{D}}}(r) - r^{-1}_{\mathcal{D}(\mathbf{v})}) \sum_{\mathbf{u} \in \mathcal{D}} f(u)}{\max_{r \in r_{\mathcal{D}}}(r) f(\mathbf{v})}$. Note that neither of the two formulas is computing Shapley values; rather, they assign a contribution to the features based on the linear weights, and the score and rank. This implies that our rank QoI is the only rank QoI for Shapley values.

E ADDITIONAL DETAILS ON METHOD COMPARISONS

E.1 Fidelity

We provide more details on the Fidelity results discussed in Section 8.2.2. We compute the Fidelity of all the methods that have that property across all datasets. We use SHAP and LIME out-of-the-box so their performance is not perfect (although extremely good). We make this choice to highlight the importance of using exact Shapley values when computing local explanations, where the error in each separate explanation is important as each explanation impacts a separate person.

Table 4: Fidelity across all methods across all datasets.

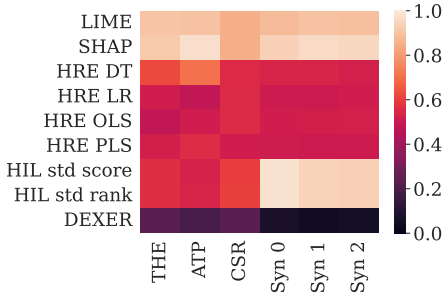
	LIME	SHAP	ShaRP		HIL
dataset	score	score	score	**rank**	score
ATP	0.98	1.00	1.00	1.00	0.14
CSR	0.95	0.99	1.00	1.00	0.85
THE	0.94	0.97	1.00	1.00	0.64
Syn 0	0.95	0.99	1.00	1.00	0.37
Syn 1	0.95	0.99	1.00	1.00	0.29
Syn 2	0.95	0.99	1.00	1.00	0.35

E.2 Agreement between Explanations

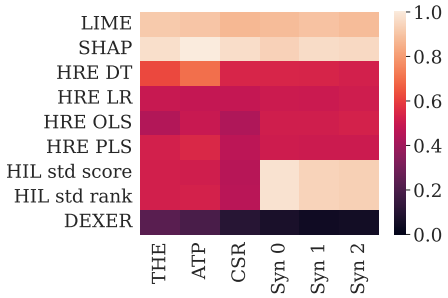
Figure 15 presents agreement between ShaRP and all other methods averaged across all datasets. We also provide the agreement per dataset in Appendix E. We use rank and score QoIs for this comparison, as they match those used by the methods we evaluate. Kendall's tau distance is computed to enable cross-method comparisons. We observe that explanations vary significantly by method, regardless of the QoI. ShaRP aligns most closely with LIME and SHAP across both rank and score QoIs. HRE, which relies on localized information, naturally differs. However, even among HRE variants, explanations remain inconsistent. The two HIL methods and the two ShaRP methods produce similar explanations despite using different QoIs, suggesting that explanation consistency depends more on the method than the QoI. In contrast, DEXER, which fits a linear regression to the ranking output and applies SHAP, differs greatly from all methods, indicating that rank cannot be effectively explained without a rank QoI.

Figure 14 provides a per-dataset visualization of the agreement between the explanations of the methods in Section **??**.

In Fig. 14a We visualize Kendall's tau explanation distance correlation of ShaRP using the rank QoI with all other methods across



(a) Method Agreement between Sharp using the rank QoI and all methods across all datasets



(b) Method Agreement between Sharp using the score QoI and all methods across all datasets

Figure 14: Method Agreement

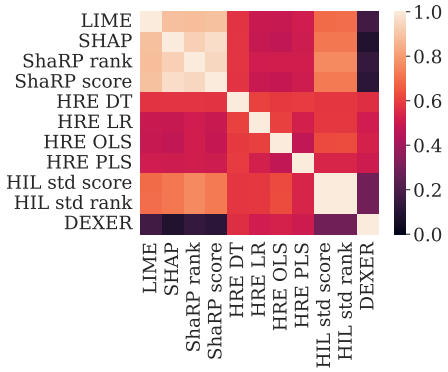


Figure 15: Method agreement averaged across all datasets

every dataset. In Fig. 14b, we plot the same result for ShaRP using the score QoI. As in the aggregated method agreement plot (Fig. 15), ShaRP is very similar to SHAP and LIME for both QoIs. As expected it is more similar for the score QoI but not identical that is perhaps because we used SHAP out-of-the-box which applies some approximation parameters for running time optimization. Similarly, ShaRP behaves similarly to what we discussed in Fig. 15 to all methods across the datasets except the HIL methods for the Synthetic datasets. We hypothesize that this is because the HIL methods are able to perform better for those datasets due to their Synthetic nature.

E.3 Sensitivity

We provide the sensitivity analysis for all methods for CSRankings in Fig. 16. In addition, to HRE LR, DEXER, LIME, SHAP, ShaRP rank, and HIL std rank that we presented in Fig. 7, we also plot HRE DT in Fig. 16a, HRE OLS in Fig. 16c, HRE PLS in Fig. 16d, ShaRP score in Fig. 16g, HIL std score in Fig. 16i. We see that all HRE methods perform similar or worse than HRE-LR, this is unsurprising as all these methods are used locally. We also see that both HIL std score and ShaRP score perform similar to SHAP, this is also expected. HIL std score and DEXER are very similar which is revealing our inability to predict the rank using the ranked output of the model. Specifically, HIL std score is assuming knowledge of the weights used by the model and uses them directly to compute the feature importance. DEXER is assuming black-box access to the ranked output only and fits a linear regression model on the ranking. Nevertheless, judging from these results, it appears that DEXER is explaining the score (and not the rank) and is learning the model weights to do so. Inadvertently, we also show that the choice of the explanation method makes a big difference to the final explanation.

To further compare ShaRP with rank QoI and HIL with Std rank, we present Figure 17. Even though both methods are appropriate for the ranking task we are examining, in this figure, we see that ShaRP with rank QoI (Figures 17a, 17c, and 17e) can capture the full range of different ranks and features, and that groups the items more successfully. HIL with Std rank cannot capture the difference of feature values for ATP (Figure 17b), or the similarly ranked items that have different feature values for the Synthetic experiment (the middle area close to the x-axis of Figure 17f. Both methods perform similarly for THE (Figures 17c and 17d).

Finally, we present an analysis of ShaRP using the score QoI and the rank QoI for the CSR dataset but for a score task (instead of rank). The goal of this analysis is to show that the sensitivity of the methods that use a score QoI is very high when we are explaining a score task. In other words, if we are trying to explain the score, then the methods that use a score-based profit function perform the best as it is fully expected.

The task we are going to explain is the score of the CSRankings scoring function. We choose this task for two reasons, first we already provided the results of the CSRankings ranking task and we can draw a direct comparison. Secondly, we have a ranking for that dataset and we can plot the methods that use the rank QoI for juxtaposition. Note that it is entirely redundant to use a rank-based QoI method in this case. In fact, it is redundant to even produce a ranking as we are asking an explainability question about the score. But we are choosing to provide this information to showcase that each explainability task needs each own profit function and the choice of the profit function makes a big difference to the final explanation. Inadvertently, we also show that the choice of the explanation method also makes a big difference to the final explanation.

In Fig. 18, we evaluate the similarity of explanations for pairs of similar items *when we attempt to explain the score*. For each pair of items, we compute three distances: (1) Euclidean distance between the explanations (x-axis); (2) distance between the *scores* (instead of rank) of the two items (y-axis); and (3) Euclidean distance between

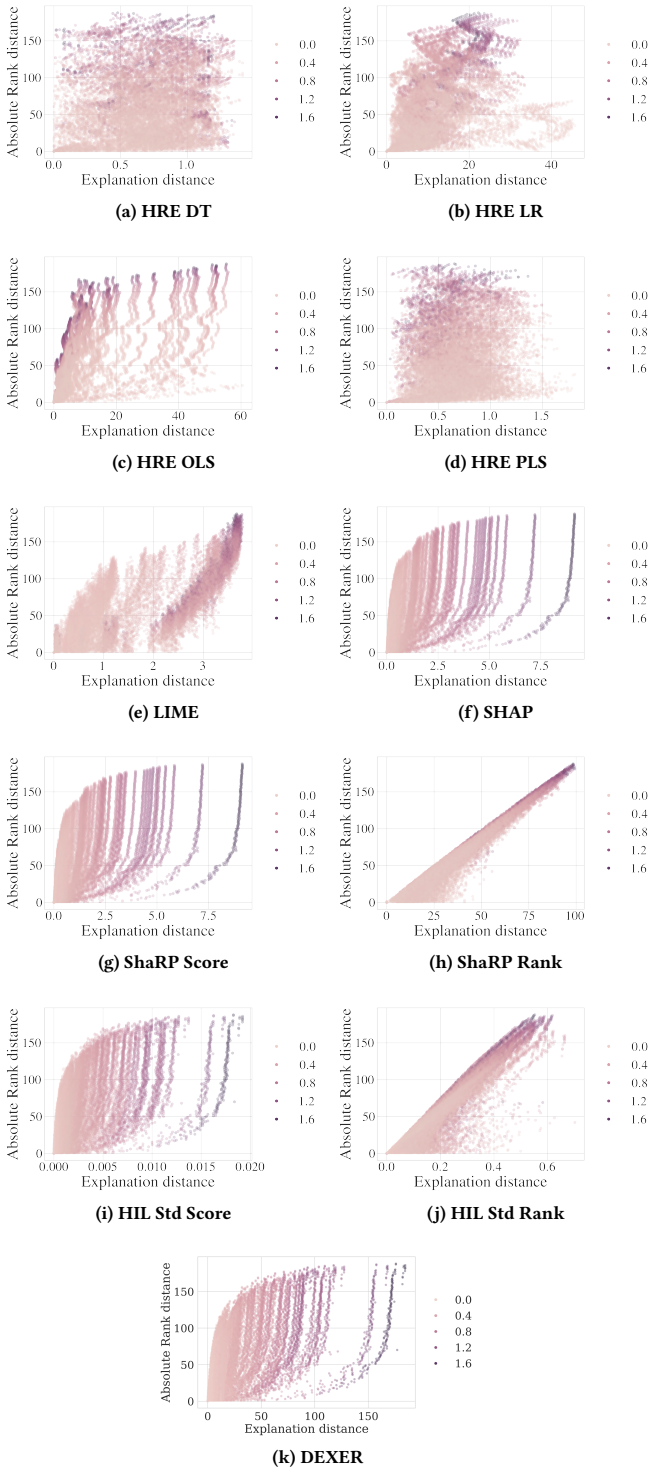
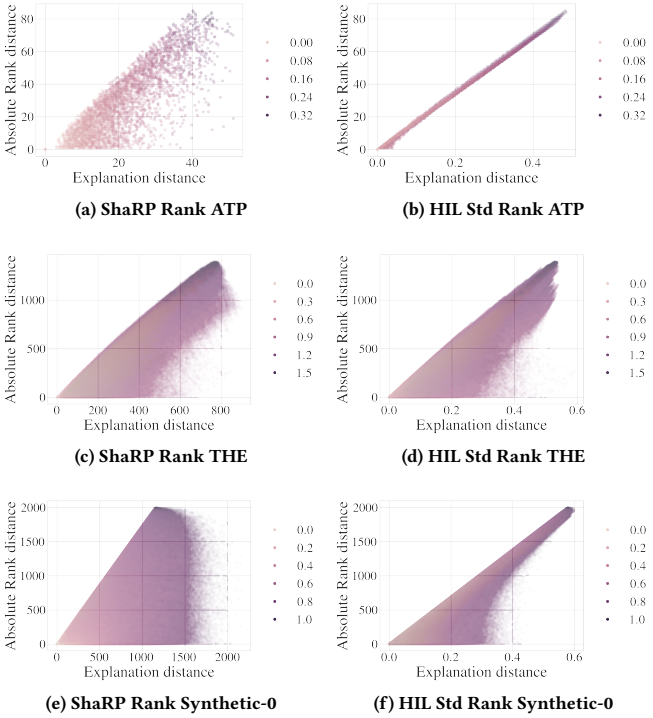


Figure 16: Sensitivity analysis for the CSRankings dataset for all methods.

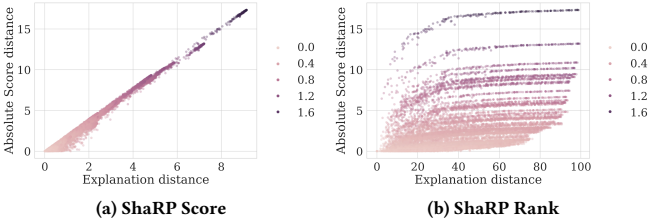


(a) ShaRP Rank ATP (b) HIL Std Rank ATP

(c) ShaRP Rank THE (d) HIL Std Rank THE

(e) ShaRP Rank Synthetic-0 (f) HIL Std Rank Synthetic-0

Figure 17: Sensitivity analysis for the ATP, THE, and Synthetic dataset 0 for the methods using the rank QoI.



(a) ShaRP Score (b) ShaRP Rank

Figure 18: Sensitivity analysis for the CSRankings dataset for all methods when the task we are trying to explain is a score task. Compared to Figure 16 we see that the methods that are using a score QoI are performing better.

the items in terms of their feature values (hue, where lighter means closer). To make the plot, we place one item (the reference item) at position (0,0) and use a scatter point for each other item (neighbor), indicating the distance in ranks and the distance of the explanations. The color of the scatter point indicates the distance between the features of the reference point and the neighbor. We then overlay the plots for all items in the dataset, so that all items are used as reference points.

Unlike Fig. 16, we now expect to see items that are both similar in terms of their features and *scored* near each other to have similar explanations. We would still expect all points to be on or near the diagonal line $y = x$, with the hue getting darker as we move away from the origin, *if their explanations successfully explain the score.*

In Fig. 18, we see that indeed the score-based methods have the desired shape we discussed in Section 8.2.1. ShaRP score is extremely similar and almost entirely fits the $y = x$ line. ShaRP rank, appears to be providing explanations that do not depend on the score distance between the items' outcomes (y-axis) or the feature distance between the items (hue) as expected.

This analysis shows how QoI selection is important when providing an explanation. The score is unable to perform well for a ranking task since it estimates the impact of each feature on the score outcome and similarly, it is completely unreasonable to use a rank QoI when explaining the score.

F ADDITIONAL RESULTS FOR QUALITATIVE ANALYSIS

In this section, we present the extended results previewed in Section 8.1.

G ADDITIONAL RESULTS ON EFFICIENCY AND APPROXIMATION

In this section, we present the extended results previewed in Section 8.3.2.

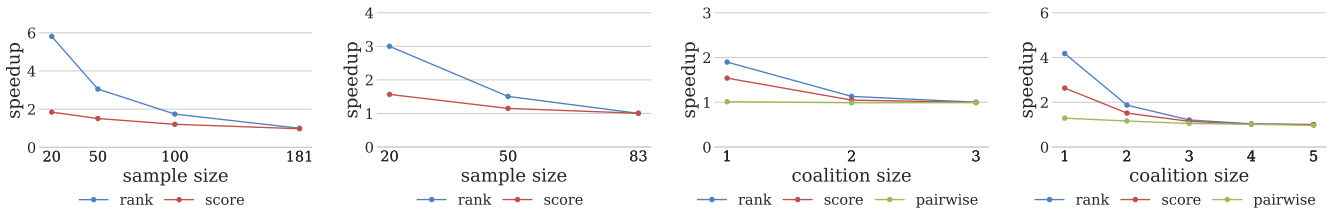
In Figure 19 we present the speedup vs. sample size, and speed-up vs max coalition size for THE, CSR, and ATP. We already presented the results for THE in Figure 8. We observe similar results but scaled down due to the dataset sizes. In Figure 20, we present the corresponding fidelity for both sample size and max coalition size. In this plot we also provide the fidelity of THE for sample size that we omitted in the main body of this work due to space constraints. We observe that fidelity is very high for all sample sizes, and almost identical to the fidelity of THE for all max coalition sizes.

In Figures 22 and 23, we present the method agreement between the approximation and the exact computation for CSRankings (CSR). We omit method agreement results for the other datasets, where ShaRP performs similarly. In 22a and 22b, we present agreement of the approximation when we vary sample size for the rank and the score QoI. We evaluate the agreement using the Jaccard Index (considering the top-2 features), Kendall's tau distance, and the Euclidean distance of the feature vectors (converted to unit vectors). Here, we see that performance is similar for both QoIs. The Jaccard index is over 0.9 for any sample size, and is the distance metric with the worst performance for both QoIs. This is worth noting as shorter explanations are often considered more interpretable [24]. Agreement is similar or higher for all QoIs when we vary maximum coalition size, see Figure 23a- 23c.

H FOCUS GROUP PROTOCOL AND RESULTS

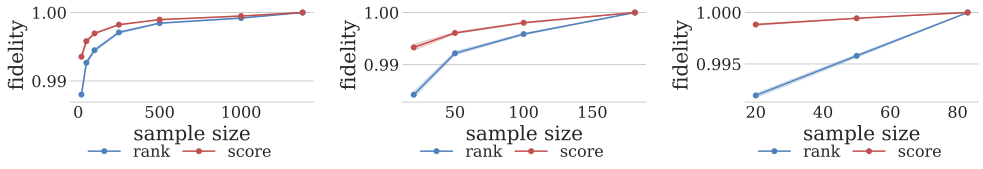
The introductory document explained CS Rankings and briefly detailed the goals of the study and the explanation method. The tasks consisted of one, two, or six explanations for CS Rankings, and the participants were asked to answer questions such as which feature contributes the most to the outcomes of one or groups of items.

To select the items presented in the study, we sampled 9 universities from CS Rankings, 3 from the top, 3 from the middle, and

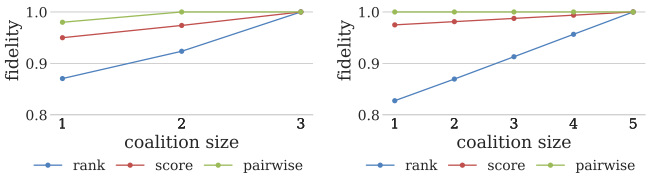


(a) Speedup vs. sample size, CSR (b) Speedup vs. sample size, ATP (c) Speedup vs. maximum coalition size, CSR (d) Speedup vs. maximum coalition size, ATP

Figure 19: Computational time performance of approximation for CSRankings (CSR), and ATP Tennis (ATP). Speedup is computed in comparison to exact computation times, reported in Table 2.

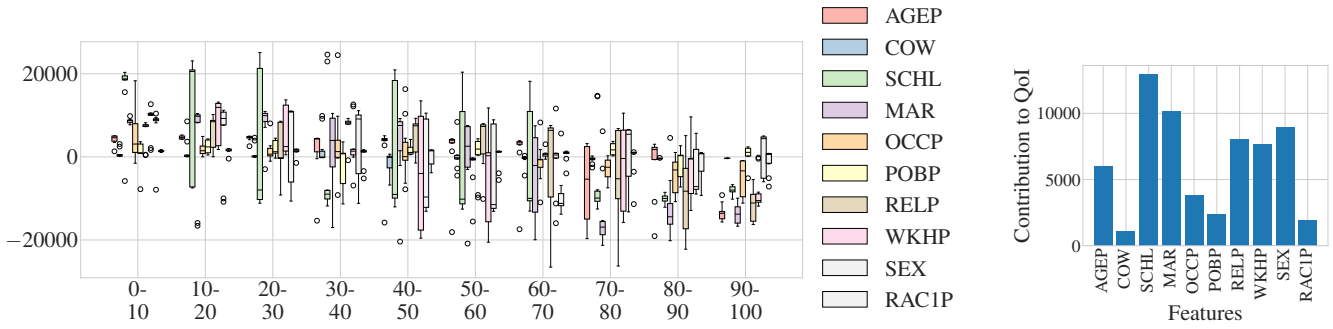


(a) Fidelity vs. maximum sample size, THE (b) Fidelity vs. maximum sample size, CSR (c) Fidelity vs. maximum sample size, ATP



(d) Fidelity vs. maximum coalition size, CSR (e) Fidelity vs. maximum coalition size, ATP

Figure 20: Fidelity of approximation for Times Higher Education (THE), CSRankings (CSR), and ATP Tennis (ATP), using different sample sizes and maximum coalition sizes.



(a) ACS strata (b) Overall

Figure 21: Feature contribution on ACS Income (Texas) to the rank QoI (a) across strata and (b) overall.

Table 5: Time experiment results over the ACS Income (Alaska) dataset. Running times for varying coalition sizes are measured using a fixed sample size of 100, while running times for varying sample sizes are measured using a fixed coalition size of 9.

		Time (cold)			Time (warm)			Fidelity		
	Approach	Pairwise	Rank	Score	Pairwise	Rank	Score	Pairwise	Rank	Score
Parameter	Value									
coalition size	1.00	1.67	1.98	0.37	1.65	1.87	0.17	1.00	0.810	0.850
	3.00	1.93	6.460	5.410	1.79	3.14	1.41	1.00	0.857	0.887
	5.00	2.34	18.57	17.70	2.01	5.82	4.13	1.00	0.904	0.924
	7.00	2.50	24.51	23.34	2.23	7.06	5.41	1.00	0.951	0.961
	9.00	2.53	24.86	23.66	2.25	7.18	5.52	1.00	0.991	0.993
sample size	20.00	—	45.10	43.35	—	12.60	11.02	—	0.994	0.996
	50.00	—	95.27	92.87	—	28.79	27.58	—	0.995	0.996
	100.00	—	160.19	154.56	—	55.80	54.54	—	0.997	0.997
	250.00	—	292.22	282.04	—	137.50	136.82	—	0.998	0.999
	500.00	—	445.00	428.96	—	271.93	270.44	—	0.999	0.999
	1000.00	—	708.37	689.55	—	542.77	536.26	—	0.999	0.999
	3348.00	—	1956.79	1960.67	—	1830.24	1816.78	—	1.000	1.000

3 from the bottom of the ranking at random. We generated explanations for all of them using our method and plotted them on the same axes so they are comparable. We created three types of tasks, first, questions of type "single" consisted of an explanation for one item and asked comprehension questions about the features importance. Questions of type "pair" asked for comparisons of the feature importance between the items. Finally, questions of type "strata" asked for comparisons between groups of items that belong to different strata. The discussion lasted 30 minutes, had prompting questions, and was overall open-ended.

Results. We found that the participants of the score-only group performed the same or worse. See Table 6. Briefly, the participants of the "rank-only" group managed to answer correctly 73% of the time, in contrast to 67% for the participants of the "score-only" group. Additionally, the participants who saw only score-based explanations were less confident in their answers, 4.15/5 and 3.90/5 measured in a 5-Likert scale, respectively. Most importantly, participants in the score-only group expressed distrust in both the ranking process and the dataset during the discussion. This is consistent with [1], which used a school admissions dataset and showed that (score-based) SHAP exhibited greater and unexplained variability in the trust of the system by users compared to other methods.

Finally, the participants in the score-only group expressed the desire to receive explanations for multiple items, instead of only one. This makes sense since a score-based explanation does not relate the item to its rank in any way. It presents the difference from its score to the mean score and is measured in score units, as opposed to the difference of the rank to the mean rank and an explanation relating each feature value to the rank. The participants said that to understand the ranking process when looking at score-based explanations, many explanations are required. This implies that a person's explanation is not sufficient to explain an item's rank in a school admissions setting for example and explanations of more people need to be shared, which could create a privacy issue.

Some relevant quotes from the participants are the following:

- "At first [for the items at the top of the ranking], the differences were so big that [the answer] was very clear, and then at the end, you know which one is better 1.05 or 1.08 [...]? So it makes you want to go back to the earlier questions and makes you question your initial impression and understanding of [the ranking]."
- "Maybe my mind started looking for some kind of [...] preconceived biases and wondering? [...] There was one figure [...] towards the end. The difference was almost imperceptible, and I kept thinking, why is one ranked few points higher than the other?"
- "I thought that the experience is successful on raising awareness and provoking critical thinking about using rankings."

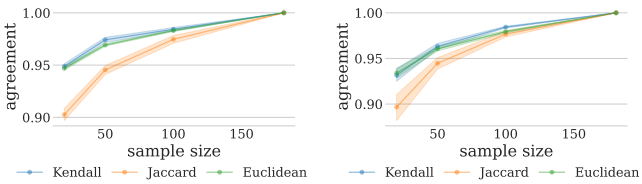
While further study is needed to investigate when and why the mistrust in the ranking process occurs and to validate the results with more participants, we believe we identified preliminary evidence that rank-based explanations are more suited to ranking tasks in both terms of understanding and trust.

I FOCUS GROUP MATERIALS

In this section, we present the materials used for the focus group described in Section 9.

Table 6: Performance in total and for each type of question. Confidence is reported on a 5 Likert scale.

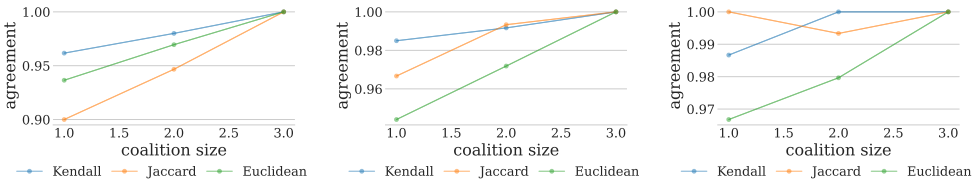
		Rank-only		Score-only	
Type	Question	% correct	Avg. Conf.	% correct	Avg. Conf.
Single	Q1	100.00%	4.43	100.00%	4.83
Single	Q2	100.00%	4.57	100.00%	4.83
Single	Q3	100.00%	4.71	100.00%	4.83
Single	Q4	85.71%	4.57	100.00%	4.67
Single	Q5	85.71%	4.14	66.67%	4.00
Single	Q6	71.43%	4.29	83.33%	4.00
Single	Q7	85.71%	4.57	83.33%	4.33
Single	Q8	85.71%	4.57	83.33%	4.67
Single	Q9	100.00%	4.29	66.67%	2.50
Single	Q10	100.00%	3.71	100.00%	4.17
Single	Q11	85.71%	4.43	83.33%	4.33
Single	Q12	85.71%	4.29	66.67%	3.00
Single	Q13	14.29%	4.29	0.00%	4.67
Single	Total	90.48%	4.38	86.11%	4.18
Pair	Q14	71.43%	4.14	0.00%	4.17
Pair	Q15	100.00%	3.86	66.67%	4.17
Pair	Q16	100.00%	3.86	83.33%	4.17
Pair	Q17	57.14%	4.00	16.67%	2.17
Pair	Q18	0.00%	4.00	50.00%	2.67
Pair	Q19	42.86%	3.71	33.33%	3.67
Pair	Total	57.14%	4.02	36.11%	3.67
Strata	Q20	14.29%	3.50	83.33%	3.67
Strata	Q21	42.86%	3.57	16.67%	3.33
Strata	Q22	71.43%	3.71	83.33%	3.00
Strata	Total	42.86%	3.63	54.17%	3.42
All	Total	72.73%	4.15	66.67%	3.90



(a) Agreement of ShaRP across different sample sizes using maximum coalition size for rank QoI

(b) Agreement of ShaRP across different sample sizes using maximum coalition size for score QoI

Figure 22: Agreement of ShaRP for CSRankings when varying the sample size.



(a) Agreement of ShaRP across different coalitions sizes using maximum sample size for rank QoI

(b) Agreement of ShaRP across different coalitions sizes using maximum sample size for score QoI

(c) Agreement of ShaRP across different coalitions sizes using maximum sample size for pairwise QoI

Figure 23: Agreement of ShaRP for CSRankings when varying the coalition size.