Allocation Schemes in Analytic Evaluation: Applicant-Centric Holistic or Attribute-Centric Segmented?

Jingyan Wang,[1,2] Carmel Baharav,[1] Nihar B. Shah,[1] Anita Williams Woolley,[1] R Ravi[1]

[1] Carnegie Mellon University
[2] Georgia Institute of Technology

Abstract

Many applications such as hiring and university admissions involve evaluation and selection of applicants. These tasks are fundamentally difficult, and require combining evidence from multiple different aspects (what we term "attributes"). In these applications, the number of applicants is often large, and a common practice is to assign the task to multiple evaluators in a distributed fashion. Specifically, in the often-used **holistic** allocation, each evaluator is assigned a subset of the applicants, and is asked to assess all relevant information for their assigned applicants. However, such an evaluation process is subject to issues such as miscalibration (evaluators see only a small fraction of the applicants and may not get a good sense of relative quality), and discrimination (evaluators are influenced by irrelevant information about the applicants). We identify that such attribute-based evaluation allows alternative allocation schemes. Specifically, we consider assigning each evaluator more applicants but fewer attributes per applicant, termed **segmented** allocation. We compare segmented allocation to holistic allocation on several dimensions via theoretical and experimental methods. We establish various tradeoffs between these two approaches, and identify conditions under which one approach results in more accurate evaluation than the other.

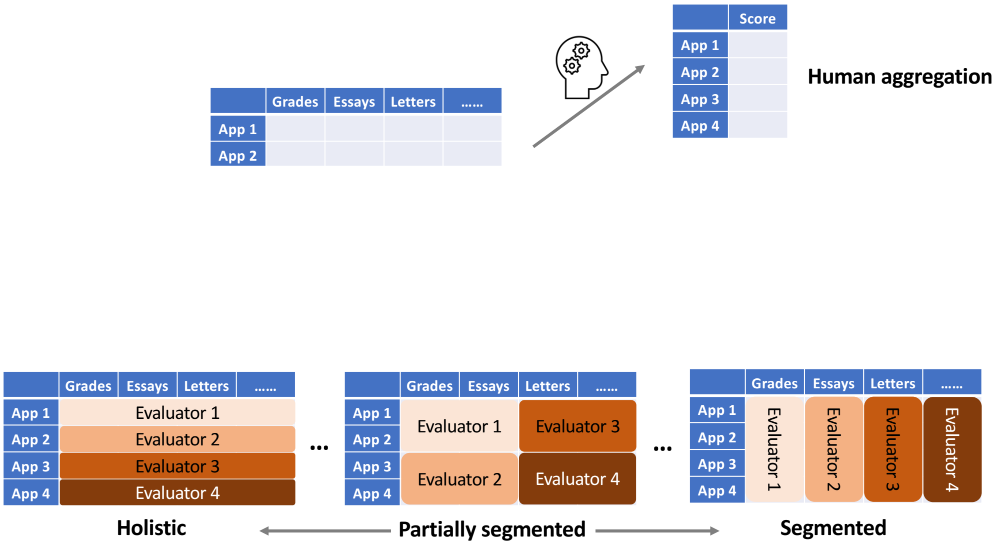
1 Introduction

Evaluation and selection are two essential functions that play a critical role in almost every organization as they determine who joins the organization, who remains, and the resulting organizational performance. However, they can also be a significant source of errors, leading to bad selection decisions and potentially limiting opportunities for certain groups. In the past, concerns about inaccurate or biased selection decisions have led to recommendations for the use of structured processes, such as structured job interviews, so that they are consistently conducted and fair to all applicants (Schmidt and Hunter 1998).

At the other end of the spectrum involving lower-stakes selection problems, distribution and automation of the evaluation task has become popular. Over the last few decades, developments in online collaboration and decision making have demonstrated the benefits of using the "crowd" for many decisions (Surowiecki 2005), opening up new possibilities for conducting evaluation and selection in a more efficient, accurate, and potentially less biased manner. For some decisions, crowd-based processes produce better results when aggregating human inputs by algorithms. However, there have also been ample examples of less effective decisions arrived at by crowds (Hube, Fetahu, and Gadiraju 2019). In reviewing the typical approaches to these crowd-based evaluations and decisions, it appears that the way they are structured varies considerably in terms of the kinds of information reviewed by evaluators when making decisions (Draws et al. 2021). We investigate the intricacies when taking this idea of distributed judgment back to the high-stakes regime, and study how the structure of an evaluation process influences the quality of decisions.

Figure 1(a) summarizes the design choices involved in an evaluation procedure. In this work, we focus on ***analytic*** evaluation, where the evaluation of an applicant (e.g., a job candidate) is decomposed into a pre-defined set of attributes. Analytic evaluation is commonly used in hiring, admissions and grading. For example, in admissions, the attributes may include the student's school GPA, essay quality, and the strength of recommendations letters. On the other hand, in ***non-analytic*** evaluation, the evaluator is not required to separately examine individual attributes. Instead, it is sufficient to provide an overall score to each applicant. While not defining attributes in the non-analytic regime offers evaluators the freedom to comprehensively think about all possible aspects of the applicants, the lack of structure may cause the evaluators to overly rely on their general impression, leading to inconsistency and inaccuracy as compared to the analytic approach (Jönsson, Balan, and Hartell 2021). Hence, analytic evaluation is our regime of interest.

Another design choice is the method to aggregate attributes to derive an overall score for each applicant. We consider ***exogenous*** aggregation, where attributes are aggregated using pre-defined (the simplest example is to take a mean, or a weighted mean, of all attributes), or algorithmically learned (Noothigattu, Shah, and Procaccia 2021) rules. On the other hand, ***human*** aggregation means that after evaluating individual attributes, the evaluator additionally provides a final score by combining the attributes in some sensible way of the evaluator's choice. Although human aggregation hypothetically provides more flexibility, simple ex-



(b)

Figure 1: An illustration of the difference between non-analytic and analytic evaluation approaches (top panel), and the spectrum of holistic vs. segmented allocation (bottom panel).

ogenous aggregation rules often turn out to be no less accurate, or even outperform human aggregation (Kahneman, Sibony, and Sunstein 2021). All in all, the issues with the non-analytic approach or human aggregation reveal that the supremacy of human reasoning is overestimated: "People trust that the complex characteristics of applicants can be best assessed by a sensitive, equally complex human being. This does not stand up to scientific scrutiny" (Highhouse 2008). Hence, our regime of interest is **analytic evaluation under exogenous aggregation**.

A fundamental question in structuring the evaluation process is how to allocate applicants and attributes to evaluators. There are two basic approaches which are likely to lead to different outcomes (see Figure 1(b)). First, in ***holistic*** allocation, evaluators are asked to review and assess all attributes about each applicant. As shown in Figure 1(b), if we assume each evaluator is represented by a rectangle of a fixed area (workload), then holistic allocation entails rectangles of the longest width (number of attributes), and therefore the smallest height (number of applicants). In ***segmented*** allocation, if we hold the workload constant, each evaluator reviews one or a few attributes for a larger number of applicants (see the right end of Figure 1(b)).

Holistic allocation is quite common in organizational hiring processes as well as in academic admissions (De Los Reyes and Uddin 2021), where people feel that a more complete understanding of an applicant or the nuances of human judgment improves the quality of decisions. Segmented allocation is more likely to be used when the attributes are considered relatively independent of one another. One example where a segmented approach is common is the grading of assignments in educational settings, where different instructors may grade different questions since performance on one is not viewed as relevant to the evaluation of another.

Here, however, we raise the question of *whether holistic allocation is as effective as often assumed, or might it be the case that segmented allocation would result in better decisions*? Certainly, for segmented allocation, it is necessary that the attributes being evaluated are separable enough that independent evaluations of them are feasible, which holds true in many instances. In these instances, we argue that segmented allocation could result in better decision quality, at least under certain conditions.

We provide a brief review of the relevant literature and outline a framework describing the key difficulties associated with evaluation that have been identified in extant research, including **calibration** of evaluators, **efficiency** with which evaluation is conducted, and the degree of **bias** in the resulting decisions. In presenting our framework, we also delineate specific conditions under which we expect that holistic or segmented allocation leads to better decision outcomes.

We employ a mixed-method study combining modeling, simulation, crowdsourcing experiments and theoretical analysis to explore the conditions under which holistic or segmented allocation performs better in terms of calibration, efficiency and fairness. In brief, we find that segmented allocation provides benefits for evaluator's calibration accuracy, whereas holistic allocation leads to greater efficiency in carrying out evaluations. In terms of mitigating bias, we observe mixed results depending on specific conditions of the application under consideration. Taken together, our work integrates a few lines of research with implications for the quality of evaluation decisions in a variety of different environments, and provides guidance to system designers for determining the evaluation structure that works best in their context.

We discuss a few key differences between our work and prior literature. At a high level, holistic and segmented allocations concern about decomposing and distributing a complex task into smaller parts. In crowdsourcing, complex tasks, such as creating animated movies, making course videos or writing articles, are broken down into parts in a similar fashion. Different crowdsourcing workers complete different parts, and then their work is computationally or manually put together (Kittur et al. 2011; Retelny et al. 2014; Cheng et al. 2015). While these works measure the quality of the completed work, such as by rating the written articles by professional journalists, we focus on more concrete and quantitative impacts of such decomposition in an evaluation and selection context, where some considerations we study such as fairness naturally arise. Another important application of human computation is peer review, where there is a large body of work on assigning reviewers to papers (Shah 2022, Chapters 3 and 4), with a focus on finding the assignments that maximize the expertise of reviewers assigned to the papers or mitigating undesirable behavior by reviewers. In our work, we primarily consider applications where the work for evaluating individual attributes can be effectively decomposed. On the other hand, in peer review, the task usually cannot be readily decomposed, and reviewers are required to read the entire paper. We discuss more related work when we formally introduce specific dimensions in Section 2.

All experiments conducted in this paper were approved by the Institutional Review Board (IRB) at Carnegie Mellon University. The crowdsourcing data, the user interface, as well as all code to reproduce our results is available at https://github.com/jingyanw/segmented-vs-holistic.

2 Theoretical Background

In theorizing the conditions under which holistic or segmented allocation leads to better decision making, we identify a few key difficulties from extant research, including *calibration* of evaluators, *efficiency* with which evaluation is conducted, and the degree of *bias* mitigation. Along these key dimensions, we present six hypotheses, study three of them in detail using theory, simulation and experiments, and leave the remaining three hypotheses for future work.

2.1 Calibration

In the context of evaluation, we use "calibration" to refer to the ability of evaluators to apply consistent criteria in assessing applicants, such that the evaluation accurately reflects each applicant's quality relative to the entire pool (Osborne 1991). Note that if an evaluator is able to perfectly identify the placement of each applicant with respect to all others under consideration, then the evaluator identifies a perfect ranking of all applicants. However, the following reasons hinder the evaluator's ability.

Lack of information about the population. In many situations, evaluators lack complete information about the full range of quality represented by applicants in the pool, and thus are not able to calibrate their assessment perfectly. Eliciting ordinal data such as pairwise comparisons or rankings (Shah et al. 2018) helps mitigate miscalibration. Nevertheless, ratings have their own benefits (Wang and Shah 2019) and ratings of some form are widely used in practice to compare applicants assessed by different evaluators. For instance, the applicants are placed in categories such as {definitely admit, maybe admit, waitlist, do not admit} in admissions, and employees are placed in categories such as {above average, below average} in performance evaluation (Goffin and Olson 2011).

We expect that issues related to evaluator calibration are among the major drawbacks of holistic allocation. In holistic allocation, each evaluator assesses all attributes for each applicant they are assigned. With the exception of very small applicant pools, this necessitates that each evaluator only sees a small subset of the entire pool. By contrast, in segmented allocation, each evaluator sees a much larger set of the pool, perhaps even the entire set of scores in the pool for their assigned attributes. Therefore, we expect that segmented allocation has an advantage of enhancing evaluator calibration.

Although it seems intuitive that evaluating more applicants improves calibration, it is unclear if it actually manifests in practice. To see a counter-argument, consider the following pair of scenarios. In the first scenario, the evaluator reviews 5 applicants whereas in the second scenario, the evaluator reviews 20 applicants. One may expect that when evaluating the last few applicants in the second scenario, the evaluator has already seen many more applications than in the first scenario. However, the evaluator may only be able to keep in mind 5 or fewer applicants when evaluating any other applicant, in which case their calibration in both scenarios will be comparable.

Hypothesis 1 (Studied in Section 4.1)**.** *For each individual attribute, segmented allocation, in which each evaluator has access to more applicants, leads to better calibration.*

Ordering effect. The lack of information about the population suggests that calibration depends on the total number of applicants assigned to evaluators. Calibration may further vary as a function of the ordering that these applicants are evaluated. One reason for such variation is the bounded rationality of people, such as the cognitive effects of primacy and recency (Page and Page 2010), assimilation and contrast (Damisch, Mussweiler, and Plessner 2006), and generosity-erosion (Vives et al. 2021). A second reason for such variation is that evaluators gradually adapt their calibration as they evaluate each applicant along the way: When an evaluator rates the 5^{th} applicant, their grading scale is based on the first 5 applicants seen so far, but by the time the evaluator moves to rate the 100^{th} applicant, they have acquired much more information for calibration from the 100 applicants compared to when they rate the 5^{th} applicant.

Such ordering effect can be mitigated in segmented allocation: Since the attributes of the applicants are assigned to different evaluators, the ordering can be shuffled so that each evaluator sees a different ordering, thereby "averaging out" the effect of ordering when the scores from these evaluators are aggregated.

Hypothesis 2. *Segmented allocation, in which the ordering* *of the applicants can be shuffled independently for each attribute, leads to better calibration compared to holistic allocation, under which all attributes are evaluated under one ordering by design.*

2.2 Efficiency

Selection and evaluation processes can also be resource-intensive and time-consuming. One reason of why quality might suffer is the basic human tendency to "satisfice" (Hilbert 2012), particularly when workload is high. Consequently, we contend that another important element to consider in evaluating the relative benefits of different allocation schemes is the impact on efficiency. This pertains to how quickly evaluators make their assessments, but also the degree to which an allocation scheme affords evaluators an opportunity to find shortcuts to adaptively allocate their effort.

Adaptively allocating effort. The goal of many evaluation and selection processes is to identify the best subset of applicants from the available pool. In holistic allocation, if a particular applicant is clearly below the threshold on a subset of the attributes, the evaluator may conclude that the applicant will not be selected, without scrutinizing the remaining attributes or giving a precise score to the applicant. In addition, evaluators may use signals, such as red flags in recommendation letters in academic admissions, to draw a preliminary conclusion which they quickly confirm or deny with a cursory review of the remaining information. The evaluators also enjoy the flexibility to adaptively choose which attribute to review next based on the attributes already reviewed.

In contrast, adaptive strategies are more challenging to implement in segmented allocation, because the evaluation task is typically allocated in parallel to the evaluators. That said, within segmented allocation, the system could employ a filtering rule for certain attributes *before* assigning applicants to evaluators. For example, in academic admissions, threshold values for standardized test scores and GPAs are often used as preliminary filters to eliminate some applicants from further consideration. However, there are concerns such that standardized test scores are themselves biased against certain groups of applicants. Another remedy is to decompose the evaluation task into multiple rounds, where applicants are filtered in between rounds. However, having multiple rounds adds logistical complexity to the evaluation procedure, and may also require more time to complete the evaluation process.

We hypothesize that in holistic allocation, evaluators can reap the adaptive benefits of efficiency without significantly sacrificing accuracy. Furthermore, we postulate that the gain is more prominent when the attributes being evaluated are correlated with one another: Screening applicants primarily based on the assessment of one attribute will be less likely to lead to errors in the overall assessment, when attributes are highly correlated than when they are only weakly correlated or independent.

Hypothesis 3 (Studied in Section 4.2)**.** *Holistic allocation results in more efficiency in evaluation without significantly reducing accuracy, when the attributes being assessed are highly correlated and thus can be used as proxies or screening tools for one another.*

Switching costs. In holistic allocation, the evaluator primarily switches between different attributes, whereas in segmented allocation, the evaluator primarily switches between applicants. Whether switching between applicants or attributes involves greater effort depends on the user interface, where the evaluator accesses applicant information by, for example, navigating through directories or downloading applicant files. A system for admissions, for example, may be designed such that more clicks are needed to access different applicants than different attributes within the same applicant. In this case, holistic allocation may incur lower switching costs than segmented allocation. However, in addition to the operational cost incurred by the user interface, another consideration relates to the cognitive load of switching between different types of information. For example, in assessing applicants for admissions, if an evaluator operating in holistic allocation has to shift from reviewing transcripts and test scores to evaluating essays and recommendation letters, the time and cognitive effort involved in this transition between attributes may outweigh the savings gained from the user interface. Consequently, whether holistic or segmented allocation leads to greater efficiency as a result of reduced switching costs depends on the user interface and the similarity in reasoning about different attributes.

Hypothesis 4. *(a) Holistic allocation results in more efficiency than segmented allocation, when transitioning from one applicant to another requires more time or clicks than transitioning between attributes of the same applicant.*
(b) Segmented allocation results in more efficiency than holistic allocation, when transitioning from one attribute to another requires high cognitive effort due to the level of variation in the data and assessment process, taking more time than transitioning between applicants for the same attribute.

We remark that the user interface should be designed to support the chosen allocation scheme. Specifically, if segmented allocation is used, then the interface should be constructed so that the switching cost between applicants for the same attribute should be made as low as possible.

2.3 Mitigating Bias

A major concern that regularly arises in evaluation and selection processes is that of bias. Researchers consider a decision to be biased when there is deviation from what is normatively predicted by classical probability and utility theory to be the optimal outcome based on the information or options available (Hilbert 2012). Bias in decision making is a widely-studied topic in a number of fields, as it has substantial implications not only for evaluation and selection decisions, but also for many other high-stakes applications such as medical diagnosis, crime prevention, and financial performance, to name just a few (Saposnik et al. 2016; Costa et al. 2017; Kovera 2019).

One type of biases of particular concern for evaluation and selection is those that result in systematic discrimination against certain groups on the basis of information that is irrelevant or inappropriate for assessment (Bertrand and Mullainathan 2004; Moss-Racusin et al. 2012; Tomkins, Zhang, and Heavlin 2017; Shah 2022, Section 7). Many biases operate on a subconscious level (Greenwald, Nosek, and Banaji 2003) and thus affect evaluations even when the evaluator intends to be fair. Consequently, the common recommendations include limiting subjective human judgment by using objective measures wherever possible, or when humans are making subjective assessments, ensuring that those are guided by specific outcome-relevant criteria and structured for consistent application to each applicant (Campion, Pursell, and Brown 1988; Pogrebtsova, Luta, and Hausdorf 2020).

We propose that the allocation scheme also has an impact on mitigating bias. Specifically, we anticipate that holistic and segmented allocations affect outcomes by limiting the impact of highly biased evaluators on overall decision accuracy, and restricting access to biasing information.

Reducing the impact of biased evaluators. It is likely that different evaluators are biased to different extents. When some evaluators are biased and some are not (or less so), holistic and segmented allocations are likely to lead to different types of impact. In holistic allocation, any particular applicant has a certain probability of being assigned a biased evaluator (depending on the fraction of biased evaluators). Consequently, a subset of the applicants are be highly affected by biased decisions, while the rest of the applicants are not. By contrast, in segmented allocation, the probability that all attributes of a particular applicant are assessed by highly biased evaluators becomes lower; however, it is more likely that each applicant receives some assessment from at least one biased evaluator, compared to holistic allocation.

Hypothesis 5 (Studied in Section 4.3)**.** *Compared to holistic allocation, segmented allocation better mitigates the impact of biased evaluators on the accuracy of the applicant evaluation, by reducing the chances that all attributes of any particular applicant are evaluated by biased evaluators.*

Restricting access to biasing information. Arguably, many interventions that have been made over the last several decades in traditional evaluation and selection processes are focused on limiting evaluators' access to biasing information. One famous example comes from symphony orchestras as they made efforts to incorporate more female musicians in the 1970s and 1980s (Goldin and Rouse 2000). Initial diversity efforts yielded limited progress, even when many orchestras conducted auditions using screens to block evaluators' view of the candidates. However, one observant evaluator noted the difference in the sound on the wooden stage floor as the musicians entered for their audition, particularly the distinct sound of the high heels worn by the female musicians in contrast to the flat sounds made by most men's dress shoes. Consequently, a number of groups began using a carpeted walkway in addition to the screen, which resulted in a sudden increase in the number of women invited to join (Goldin and Rouse 2000). In generalizing this idea to the context of evaluation and selection, we hypothesize that segmented allocation mitigates bias by limiting access to information about irrelevant and potentially biasing attributes. For example, an evaluator could be asked to evaluate the research statements of graduate school applicants without access to any other information about the applicants, substantially limiting the possibility of bias.

Hypothesis 6. *Segmented allocation helps mitigate the impact of bias compared to holistic allocation, as a result of limiting evaluators' access to biasing information when they assess individual attributes of the applicants.*

3 Modeling Framework

We describe the mathematical framework used in our analysis.

Notation. We assume that there are n applicants, and each applicant has d attributes. We let $x_{ij} \in \mathbb{R}$ be the true quality of applicant $i \in [n]$ on attribute $j \in [d]$.[1] A higher value represents higher quality. When there is more than one attribute, we define the true ranking of the applicants as the ranking induced by the mean of their attribute values. The evaluation task is represented by the matrix $\{x_{ij}\}_{i\in[n],j\in[d]}$, and we divide the matrix into sub-matrices as shown in Figure 1, where each evaluator assesses a smaller sub-matrix consisting of a subset of the applicants and a subset of the attributes (where the subset is allowed to equal the entire set). For simplicity, we assume each attribute of each applicant is evaluated once, so all the sub-matrices are disjoint and collectively partition the entire matrix. We let $y_{ij} \in \mathbb{R}$ denote the score given to attribute j of applicant i by the assigned evaluator. Note that y_{ij} is often a noisy evaluation of x_{ij}.

Metric. In many evaluation and selection processes such as hiring or academic admissions, the goal is to choose a specified number of applicants of the highest quality. Therefore, the accuracy of the evaluation process is determined by the top-K accuracy in ranking. For simplicity, we consider the top-1 accuracy as studied by Kleinberg and Raghavan (2018). That is, the accuracy is 1 if the estimated ranking correctly identifies the best applicant in the true ranking, and 0 otherwise.[2] We also consider a second error of metric that is suitable for understanding the calibration of evaluators. This error represents the mean error in estimating the percentile of each applicant, described in detail in Section 4.1.

Data generation. In our simulations, we follow prior work (Kleinberg and Raghavan 2018) and generate the attribute values from the power-law distribution unless specified otherwise. The power-law distribution with parameter $\delta > 0$ is defined as $\mathbb{P}[Z \geq t] = t^{-(1+\delta)}$ supported on $t \in [1, \infty)$, where Z denotes the random variable.

We allow the attributes to be correlated, defined by a correlation parameter $\sigma \in [-1, 1]$. For any desired distribution

[1] We use the notation $[\kappa]: = \{1, 2, \ldots, \kappa\}$ for any positive integer κ.

[2] In our setup, we make sure that there exists a unique best applicant in the true ranking. If there are ties in the estimated ranking, the accuracy is computed as $1/$(number of applicants in the tie) if the true best applicant is one of the estimated applicants in the tie, and 0 otherwise.

with c.d.f. F, we define the following procedure (cf. Nelsen 2010) to generate d-dimensional correlated random variables. Let Φ denote the c.d.f. of the standard normal. For each applicant i, we first sample a vector $\boldsymbol{z}_i \in \mathbb{R}^d$ from a multinomial normal distribution as $\boldsymbol{z}_i \sim \mathcal{N}(0, (1-\sigma)I_d + \sigma \mathbf{1}_d \mathbf{1}_d^T)$ independent across the applicants $i \in [n]$. Then we compute the attribute values as $x_{ij} = F^{-1}(\Phi(z_{ij}))$. It can be verified that each x_{ij} has marginal distribution F. As special cases, when $\sigma = 1$, all attributes have identical values; when $\sigma = 0$, all attributes are independent.

4 Methods and Results

In this section, we examine our hypotheses related to calibration, efficiency and mitigating bias.

4.1 Calibration

We focus on studying the relation between calibration accuracy and the number of applicants assigned to an evaluator, as described by Hypothesis 1.

Operationalization of calibration. Formally, we define calibration as the evaluator's accuracy of estimating the ranking (or percentile) of each applicant with respect to the entire pool of all applicants. We define calibration on this relative scale for three reasons. First, the selection problem is intrinsically relative in nature, that is, we aim to select the top applicants compared to the entire pool. Second, in many applications, the evaluators are asked to report relative data. For example, evaluators may be asked to give scores on a scale of 1-5, where the criteria define the score of 1 as the applicant being the bottom 20% among all applicants, and 2 as being 20-40% among all applicants, etc. Third, social comparison theory suggests that people's reasoning has a relative nature (Festinger 1954). For example, being a "top" applicant is perceived as simply being significantly better than the rest of the applicants. For this reason, using a relative scale than an absolute scale is shown to be more effective in various judgment tasks (Goffin and Olson 2011).

Experimental setup. To isolate the impact of calibration, we make a number of design simplifications, and conduct an experiment focusing on a single attribute. We recruit 200 crowdsourcing workers on the Prolific platform. The workers are introduced to a hiring context and asked to evaluate scores of applicants. Specifically, they are told that there are 1000 applicants with scores that are integers between 0 and 300, without any distributional information about the scores. Then the workers are presented some numbers in between 200 and 300, and are asked to estimate the percentile of the scores. The workers classify each score to one of the five bins with respect to the population: 0-20%, 20-40%, 40-60%, 60-80%, and 80-100%. We choose to ask the workers to report in 5 quantized bins instead of directly reporting a number of percentile, because prior studies have shown that workers are not able to perceive fine numbers accurately due to limited processing abilities (Miller 1956; Shah et al. 2016) and therefore have higher accuracy when a small number of quantized choices are given (e.g., Lietz 2010). We have confirmed this trend by a preliminary study comparing using 5 bins versus 10 bins.

Question grouping. The workers are divided into two groups uniformly at random. Recall that there is a single attribute. In the first group, each worker is presented scores of 5 applicants (termed "5Q-group"). In the second group, each worker is presented scores of 20 applicants (termed "20Q-group"). The workers are always presented with 5 scores per page. That is, for the 20Q-group, the 20 questions are distributed across 4 pages. Neither group of workers is told the number of scores they will be presented before starting the task. The workers are required to answer all questions on a page before proceeding to the next page, though they are allowed to review and edit their answers on previous pages at any time before submission. We choose to present 5 questions per page and not inform the workers the total number of questions, to address the confounder that a worker who knows they have to do 20 questions may put less effort per question than if they knew they have to do only 5 questions.

Values of scores. Since we consider a single attribute, we use the shorthand $x_i :\ = x_{i1}$ for the true score of each applicant i. Let F be the distribution $\mathcal{N}(230, 25)$, truncated to the range of $[200, 300]$. The scores $\{x_i\}_{i \in [n]}$ in the 20Q-group are generated i.i.d. from F. We pair up workers in the 20Q-group and the 5Q-group. For the scores in the 5Q-group, we use the same values as the last 5 questions in the 20Q-group for a direct comparison. We choose this distribution for scores, because in a preliminary study where the workers are presented scores in the range of $[0, 100]$, we observe that the workers appear to have a strong uniform prior, mapping scores in $[0, 20]$ to percentile 0-20%, etc. This uniform mapping is an artifact of the experimental design that the quality under evaluation is real-valued. In more realistic situations, such a simplified mapping, say from applicants' interview performance to scores, does not exist. We therefore choose a range that is not $[0, 100]$ so that the workers do not rely on such priors.

Experimental Results. We record the worker calibration measured by their accuracy in estimating the percentile bins. Formally, let ω be the function mapping the percentile 0-20%, 20-40%, 40-60%, 60-80% and 80-100% to the bins $1, 2, 3, 4$ and 5, respectively. For a single worker, let $y_i \in [5]$ be the bin reported for applicant i. Then the absolute error between the true bin and the reported bin for applicant i incurred by this worker is defined as $\left|\omega(F^{-1}(x_i)) - y_i\right|$.

For each worker, we compute their mean error over the applicants they evaluate. The workers' mean error is 1.14 ± 0.06 in the 5Q-group, and 0.84 ± 0.05 in the 20Q-group. We perform a univariate permutation test between the mean errors of workers in the 20Q-group, and those of workers in the 5Q-group, using the difference in sample means as the test statistic. We reject the null hypothesis that the errors from the two groups have the same mean (one-sided p-value < 0.01; Cohen's effect size $d = 0.52$). This result indicates that evaluation in the 20Q-group is more accurate than in the 5Q-group, confirming Hypothesis 1 that evaluators have better calibration when they see more applicants.

For the 20Q-group, we also separately compute each worker's mean error over each page of 5 questions (that is, Q1-5, Q6-10, Q11-15, Q16-20). The mean error for each

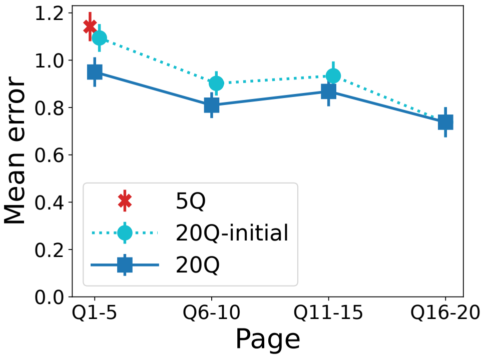


Figure 2: The mean error in estimating the percentile bins, for workers in the 5Q-group (representing holistic allocation) and the 20Q-group (representing segmented allocation). Error bars represent the standard error of the mean.

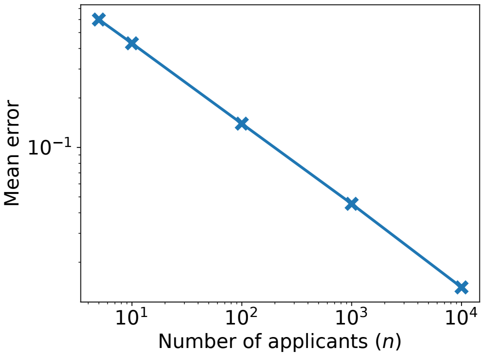


Figure 3: The mean error in calibration of a single evaluator, as a function of the number of applicants. Each point is computed over 1000 runs (error bars are too small to be visible).

page is plotted in Figure 2. For the 20Q-group, we also plot the error on each page using the answers reported right before the workers ever turn to see the next page (see the curve "20Q-initial"). The difference between the curves of the initial and final errors thus corresponds to the gain in calibration by workers correcting their answers to previous applicants by seeing applicants from later pages. First, we observe that such corrections notably decrease the error, especially for the first page. This observation further provides evidence for Hypothesis 1 by showing that workers are able to use the information they see from applicants to perform correction. Second, even after this correction, the error has a decreasing trend from earlier pages to later pages, suggesting that workers have limited abilities in performing such corrections. Specifically, in the 20Q-group, the (final) mean error for page 1 is 0.95 ± 0.06, and the (final) mean error for page 4 is 0.74 ± 0.06. We perform a univariate permutation test between the mean errors for page 1 and page 4, using the difference in sample mean as the test statistics. We reject the null hypothesis that the errors for these two pages have the same mean (one-sided p-value < 0.01; Cohen's effect size $d = 0.34$). Third, as a sanity check, we observe that for page 1, the mean error in the 20Q-initial curve is similar to the mean error of the 5Q-group. This is expected, as the workers from the two groups have strictly the same information before the workers in the 20Q-group ever turn to the second page.

We observe the same qualitative trends in a previous version of the experiment, discussed in the extended version of this paper on arXiv (Wang et al. 2022, Appendix B).

Simulations. The key observation from the crowdsourcing experiment is that seeing more applicants improves calibration. We now conduct additional simulations for a more quantitative understanding. We stick with the setting of a single attribute. We consider the following model for evaluators. When an evaluator is assigned n applicants, it assigns the lowest $\frac{n}{5}$ applicants to the bin 0-20%, followed by the next $\frac{n}{5}$ applicants to the bin 20-40%, etc. This is a natural model for evaluators, because as n goes to infinity, the mean error on the reported bins approaches 0.

We plot the mean error as a function of the number of applicants n assigned to a single evaluator in Figure 3. The error decreases as the number of applications n increases, matching the experimental result and therefore providing additional evidence supporting Hypothesis 1. We empirically observe that as the number of applicants n increases, the mean error decreases at a rate of $\frac{1}{\sqrt{n}}$.

4.2 Efficiency

We study the adaptive allocation of effort in Hypothesis 3 via simulations.

Setting. We consider $n = 200$ applicants, and for simplicity, $d = 2$ attributes assessed by two evaluators. In segmented allocation, each evaluator is assigned one attribute of all applicants. In holistic allocation, each evaluator is assigned both attributes of half of the applicants. The attribute values are generated from a power-law distribution with parameter 1, with correlation $\sigma \in [0, 1]$ between the two attributes. To isolate the efficiency aspect from calibration errors, we assume that an evaluator always reports the true value of the attributes, namely $y_{ij} = x_{ij}$ for each (i, j) pair.

According to Hypothesis 3, holistic allocation provides the opportunity for an evaluator to decide whether to evaluate the second attribute of an applicant, based on the quality of the first attribute. For simplicity, we assume that in holistic allocation, each evaluator always reviews attribute 1 of all applicants. Each evaluator then reviews attribute 2 only on the applicants who have scored high on attribute 1. Specifically, we assume that attribute 2 is only evaluated on a τ-fraction[3] of the applicants receiving the top scores on attribute 1, for a parameter $\tau \in (0, 1]$. Finally, the best applicant is selected as the one whose mean of the two attribute scores is the maximum, namely $\text{argmax}_{i\in[n]}(y_{i1} + y_{i2})$, from the applicants on which both attributes are evaluated.

[3] Selecting the top τ-fraction requires knowledge about attribute 1 of all the applicants that an evaluator is assigned. In practice, an evaluator may select the applicants whose attribute 1 exceeds a certain real-valued threshold, which approximately has the same effect.

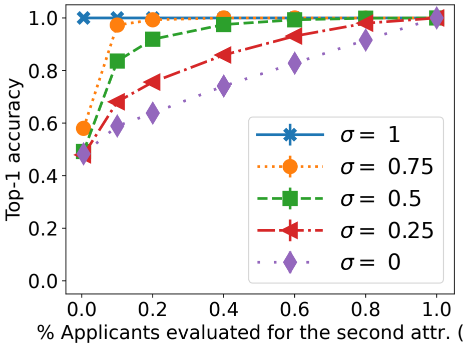


Figure 4: Top-1 accuracy for different fractions τ of the applicants evaluated for the second attribute, and various values of the correlation σ between the two attributes. Each point is computed over 1000 runs (error bars are too small to be visible).

Simulations. In Figure 4, we compute the top-1 accuracy for different fraction τ and attribute correlation σ. When the correlation is $\sigma = 1$ (see the blue curve), by definition evaluating only attribute 1 achieves perfect accuracy, and there is no need to evaluate attribute 2. When the correlation σ is relatively high, we observe that relatively small values of τ introduce a significant amount of saving in terms of the total number of attributes evaluated, while the accuracy only decreases marginally. This observation validates Hypothesis 3, and we conclude that a higher correlation between the attributes allows more saving in holistic allocation. This result points to a tradeoff between efficiency and accuracy in holistic allocation – namely, smaller τ introduces savings but also more error. The specific point to pick in this tradeoff depends on the goals of the system designer.

4.3 Mitigating Bias

To study Hypothesis 5, we present a simple model for analyzing the effect of bias reduction. We present theoretical guarantees and simulational results that characterize the regimes under which segmented allocation results in more accurate and less biased evaluations than holistic allocation. Our results provide intuition on the effect of redistributing and reducing the impact of biased evaluators.

Formulation. Recall that x_{ij} denotes the true value of applicant $i \in [n]$ on attribute $j \in [d]$. We assume that the applicants consist of two groups – advantaged and disadvantaged – where a fraction $\alpha \in [0, 1]$ of the applicants are from the disadvantaged group. We assume that a fraction $\lambda \in [0, 1]$ of the attributes are "protected". Each evaluator has an independent probability of $\gamma \in (0, 1)$ to be biased in the following sense: An unbiased evaluator reports the (noiseless) true value $y_{ij} = x_{ij}$ for any applicant i and any attribute j that they are assigned, while a biased evaluator applies a multiplicative bias factor $\beta \in [0, 1)$ to the protected attributes of the disadvantaged applicants, and reports the true value otherwise. In other words, for attribute j of applicant i, a biased evaluator reports

$$y_{ij} = \begin{cases} \beta x_{ij} & \text{if } j \text{ protected and } i \text{ disadvantaged} \\ x_{ij} & \text{otherwise.} \end{cases}$$

For ease of analysis, we consider a simple case of $d = 2$ attributes, where the correlation between the two attributes is $\sigma = 1$. That is, for each applicant i, the two attributes have identical values $x_{i1} = x_{i2}$. We hence use the shorthand x_i to denote this value. We assume that the values $\{x_i\}_{i \in [n]}$ are generated i.i.d. from a continuous[4] distribution $\mathcal{D}$ supported on $[0, \infty)$, such as the power-law distribution. Let $S_{\text{d}} \subseteq [n]$ denotes the set of αn disadvantaged applicants, and let $S_{\text{a}} \subseteq [n]$ denotes the set of $(1 - \alpha)n$ advantaged applicants. Denote the quality of the best applicant in the disadvantaged group by $x_{\text{d}}^{\max} : = \max_{i \in S_{\text{d}}} x_i$, and likewise denote $x_{\text{a}}^{\max} : = \max_{i \in S_{\text{a}}} x_i$. We compute the mean of attribute scores for each applicant, and estimate the best applicant by selecting the one with the maximum mean score. Denote the expected top-1 error under holistic and segmented allocations by e_{hol} and e_{seg} respectively, formally defined by $\mathbb{P}(\text{argmax}_{i \in [n]} x_i \neq \text{argmax}_{i \in [n]} y_{i1} + y_{i2})$, using the scores $\{y_{ij}\}$ under the two allocation schemes respectively.

Theoretical results. We focus on a simplified case of two evaluators, which as we see shortly, already illustrates the intricacy of the comparison. In this setting, holistic allocation assigns each evaluator both attributes of half of the applicants; segmented allocation assigns each evaluator one attribute of all applicants. We assume that the assignment to applicants and attributes is uniformly at random.

Theorem 1. *Let the number of attributes be $d = 2$. Let the fraction of disadvantaged applicants be $\alpha = 0.5$. Let the two attributes have identical values (that is, $x_i : = x_{i1} = x_{i2}$), sampled i.i.d. from a continuous distribution $\mathcal{D}$. Consider holistic and segmented allocations under two evaluators.*

(a) Let $\lambda = 0.5$, that is, one of the two attributes is protected. Then for any bias factor $\beta \in [0, 1)$ and any evaluator bias probability $\gamma \in (0, 1)$, segmented allocation incurs a lower error than holistic allocation, that is, $e_{seg} \leq e_{hol}$.

(b) Let $\lambda = 1$, that is, both attributes are protected. Let $\beta = 0$ (extreme downward bias against disadvantaged applicants). Then

$$e_{hol} - e_{seg} = \frac{\gamma(1-\gamma)}{2} \left[4 \cdot \mathbb{P}\left(x_d^{max} > 2x_a^{max} \right) - 1 \right] . \tag{1}$$

Hence, for any $\gamma \in (0, 1)$, segmented allocation incurs a lower error than holistic allocation, if and only if

$$\mathbb{P}\left(x_d^{max} > 2x_a^{max} \right) > 0.25. \tag{2}$$

This condition (2) is dependent on the number of applicants n and and the distribution $\mathcal{D}$, and independent of the other problem parameters. In particular, for the

[4] We consider continuous distributions for simplicity, so that the best applicant is uniquely defined with probability 1.

power-law distribution with a constant parameter δ, segmented allocation is better than holistic allocation for sufficiently large n, if and only if

$$\delta < \frac{\log(3)}{\log(2)} - 1 \approx 0.58. \tag{3}$$

The proof of this theorem is provided in the extended version of this paper on arXiv (Wang et al. 2022, Appendix A). This theorem reveals that segmented allocation is better than holistic allocation in terms of accuracy over a large range of parameters, but not always. Despite the simplified settings considered in the theorem, the result illustrates how allocating biased evaluators differently leads to changes in accuracy.

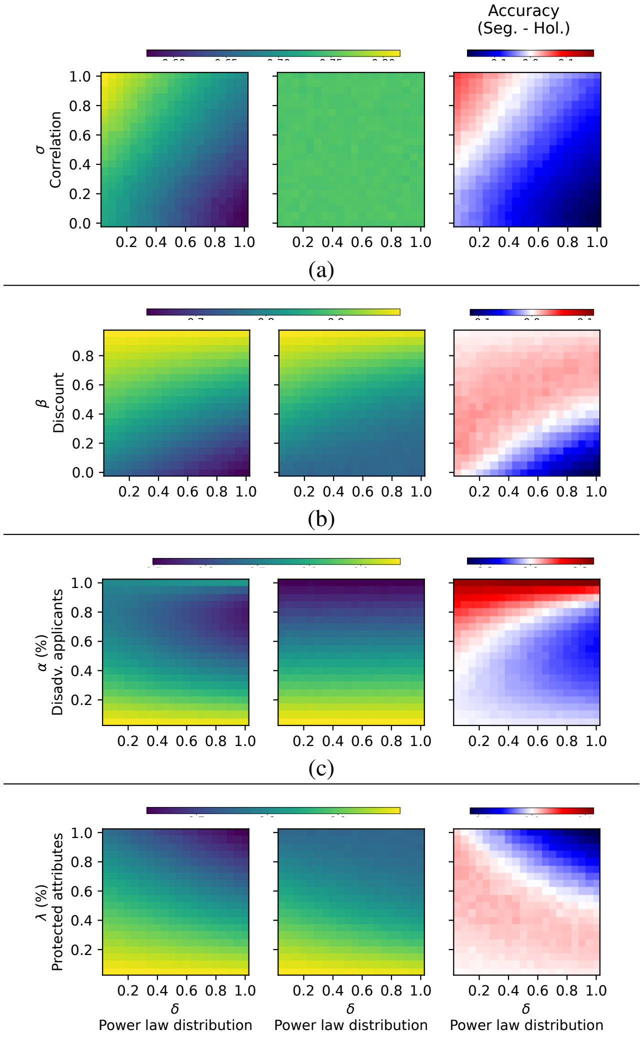
Simulations. We study the effect of the set of parameters $(\delta, \sigma, \beta, \alpha, \lambda)$ in the model. Following the assumption of Theorem 1, we consider two evaluators for simulation. The proof of Theorem 1 suggests that it suffices to consider one biased evaluator and one unbiased evaluator. We fix the number of applicants $n = 20$ and the number of attributes $d = 20$. To inspect the difference between holistic and segmented allocations, for ease of visualization, we vary two parameters at a time while keeping the other ones fixed. For consistency, one varying parameter is always δ for the power-law distribution. We set the default parameter values as $\sigma = 0.5$, $\beta = 0$, $\alpha = 0.5$ and $\lambda = 1$, when they are not chosen as the parameter to be varied. The results are shown in Figure 5 and discussed below.

Effect of power-law parameter (δ) In Figure 5(a)-(d), we observe the general trend that both segmented and holistic allocations achieve higher accuracy under smaller values of δ. A smaller δ means that the distribution has a heavier tail, so that the values of the applicants are more spread out. Hence, the best applicant has a more extremal, higher value compared to the other applicants, giving stronger signal for the evaluation process and making it easier. Two exceptions to this general trend are holistic allocation in Figure 5(a) and 5(c), where the accuracy is independent of δ. In these two cases, we have $\lambda = 1$ and $\beta = 0$. Hence, when a biased evaluator is assigned a disadvantaged applicant in holistic allocation, all attributes ($\lambda = 1$) are discounted to zero ($\beta = 0$), making it impossible for disadvantaged applicants to be identified as the best regardless of their values, and thus the accuracy is independent of δ.

Effect of correlation (σ): Figure 5(a) In holistic allocation, for the same reason that the accuracy is independent of δ as previously explained, the accuracy is also independent of σ. In segmented allocation, we observe that a higher correlation leads to a higher accuracy. This is because a higher (positive) correlation strengthens the signal for applicants. For example, consider the extreme case of $\sigma = 1$. Then the same attribute value is replicated d times for each applicant, improving robustness against randomness in the evaluation process due to bias.

Comparing segmented and holistic allocations, we observe that segmented allocation performs better when σ is high (more correlation) and δ is small (heavy tail in the distribution). The tradeoff between the two allocation schemes arises, because segmented allocation always discriminates disadvantaged applicants but to a lesser extent, whereas holistic allocation discriminates disadvantaged applicants less often but to a greater extent. When the correlation σ between the attributes is high, the gain from only discriminating a fraction of the attributes (as supposed to all attributes) is more significant.

Finally, note that in Theorem 1 we set the correlation as $\sigma = 1$. Hence, the setting of Theorem 1(b) corresponds to the top most matrix row in Figure 5(a). We observe that the sign of the comparison between the two schemes is consistent with the theoretical result, with a change-point at $\delta \approx 0.6$ in the right panel of Figure 5(a).



(a)

(b)

(c)

(d)

Figure 5: [Best viewed in color.] Comparison between holistic and segmented allocations, under different parameter values for δ and $(\sigma, \beta, \alpha, \lambda)$. The plots depict the accuracy of segmented (left) and holistic (middle) allocations, and their difference (right). For each choice of the parameters, the accuracy for the two allocation schemes is respectively computed over 50000 runs.

Effect of bias factor (β): Figure 5(b) We observe that both allocation schemes have a higher accuracy when the value of β is large. This is natural because a larger β corresponds to less discrimination by the biased evaluators. Comparing the two schemes, segmented allocation is more advantageous when β is larger. We reason that when β is small (more discrimination), the effect when a disadvantaged applicant is discriminated is very detrimental for the applicant (either on one attribute for segmented allocation, or both attributes for holistic allocation). Hence, holistic allocation performs better because the probability that a disadvantaged applicant is discriminated (to any extent) is smaller. On the other hand, when β is large (less discrimination), the advantage of segmented allocation of discounting only on one attribute becomes more significant, as supposed to discounting both attributes in holistic allocation.

Effect of fraction of disadvantaged applicants (α): Figure 5(c) We observe that segmented allocation performs better in general when α is large. To reason about this effect, let us first think about the extreme case when $\alpha = 1$. In this case, segmented allocation is better because it gives consistent treatment to all applicants. Namely, the biased evaluator discounts one attribute of all applicants. On the other hand, holistic allocation creates discrepancy between applicants, because only the disadvantaged applicants assigned to the biased evaluator are discounted. Moreover, note that the performance of segmented allocation is not monotonic in α: For larger values of δ, segmented allocation has the lowest accuracy when a large fraction, but not all applicants are disadvantaged. This non-monotonicity of segmented allocation leads to the non-monotonicity in comparing the two schemes in the right panel of Figure 5(c).

Effect of fraction of protected attributes (λ): Figure 5(d) We observe that segmented allocation performs better when λ is small. In this case, there is less discrimination in both allocations: Segmented allocation decreases the probability that a biased evaluator is assigned a protected attribute, whereas holistic allocation decreases the impact of a biased evaluator on an applicant. Our empirical observation aligns with the theoretical results: Comparing part (a) and part (b) of Theorem 1 also suggests that segmented allocation performs better for smaller values of λ.

In summary, there is a tradeoff where more segmentation means that the disadvantaged applicants are more likely to be consistently discriminated, but to a lesser extent; the parameters tip this tradeoff in different manners. We conclude that Hypothesis 5 does not capture the complete picture, as the benefit of segmented allocation depends on the specific values of the parameters.

5 Discussion

In this work, we consider using segmented allocation as an alternative to the conventional holistic allocation, for applications such as hiring and admissions. We provide detailed discussions comparing the two allocation schemes, and present theoretical and experimental results focused on three aspects: calibration, efficiency and fairness. These results indicate the potential improvement by segmented allocation on calibration, while also suggesting that holistic allocation has potential benefits on efficiency. The two allocation schemes also distribute evaluators differently that lead to different impacts in terms of fairness. These results together suggest a tradeoff between holistic and segmented allocations (and the spectrum in between). The tradeoff and the choice of which allocation to use depends on the characteristics of specific applications and which dimensions are prioritized by the system designer.

Immediate open problems include validating the remaining three hypotheses that are not analyzed in this paper, and extending the theoretical and simulation results to more general scenarios to improve our understanding of the bias considerations in Section 4.3. For example, if each attribute of each application is evaluated by many evaluators, then it is natural to expect that the bias is averaged out more evenly across evaluators, and the discrepancy between holistic and segmented allocations becomes less prominent. There are also various other considerations, as well as open problems:

- Segmented allocation requires grouping of attributes, and the system designer needs to do this grouping appropriately. For example, in the case of admissions, one may group test scores and GPAs as one attribute called "scholarly performance". In order to provide appropriate context to evaluators, one may also need to provide the same attributes to multiple evaluators.
- In addition to grouping the attributes, it is also possible to group the applicants. We have assumed that the applicants are distributed to evaluators uniformly at random. In reality, evaluators may have different expertise that make them more suitable to review a particular subset of the applicants. For example, in admissions, evaluators from the same educational background as the applicants may be more familiar with interpreting the schools and the GPAs.
- We have assumed for simplicity that the final score is computed by taking the mean over all attribute scores. In practice, we may want to use different weights for different attributes, or even use non-linear functions. In some cases, the aggregation function may not be precisely provided by the system designer, but needs to be learned from past data. This problem of learning the aggregation function for evaluation has been studied in the specific context of peer review (Noothigattu, Shah, and Procaccia 2021), and it is of interest to extend such results to more general applications.
- This work discusses a spectrum of choices in terms of the number of attributes and applicants assigned to each evaluator. An open problem of interest is to establish the optimal point(s) on this holistic-segmented spectrum theoretically and practically for any given specification of the applications and desiderata.

Acknowledgments. We thank Lily Laredo for inputs on writing the worker instructions for the crowdsourcing experiments. This work was supported by the CMU Block Center and NSF CAREER 1942124.

References

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4): 991–1013.

Campion, M. A.; Pursell, E. D.; and Brown, B. K. 1988. Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology*, 41(1): 25–42.

Cheng, J.; Teevan, J.; Iqbal, S. T.; and Bernstein, M. S. 2015. Break it down: A comparison of macro- and microtasks. In *Proceedings of the Conference on Human Factors in Computing Systems*, CHI '15, 4061–4064.

Costa, D. F.; de Melo Carvalho, F.; de Melo Moreira, B. C.; and do Prado, J. W. 2017. Bibliometric analysis on the association between behavioral finance and decision making with cognitive biases such as overconfidence, anchoring effect and confirmation bias. *Scientometrics*, 111(3): 1775–1799.

Damisch, L.; Mussweiler, T.; and Plessner, H. 2006. Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, 12(3): 166–78.

De Los Reyes, A.; and Uddin, L. Q. 2021. Revising evaluation metrics for graduate admissions and faculty advancement to dismantle privilege. *Nature Neuroscience*, 24(6): 755–758.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, HCOMP 2021, 48–59.

Festinger, L. 1954. A Theory of social comparison processes. *Human Relations*, 7(2): 117–140.

Goffin, R. D.; and Olson, J. M. 2011. Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, 6(1): 48–60.

Goldin, C.; and Rouse, C. 2000. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4): 715–741.

Greenwald, A. G.; Nosek, B. A.; and Banaji, M. R. 2003. Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2): 197.

Highhouse, S. 2008. Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, 1: 333 – 342.

Hilbert, M. 2012. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, 138(2): 211.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the Conference on Human Factors in Computing Systems*, CHI '19, 1–12.

Jönsson, A.; Balan, A.; and Hartell, E. 2021. Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice*, 28(3): 212–227.

Kahneman, D.; Sibony, O.; and Sunstein, C. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Kittur, A.; Smus, B.; Khamkar, S.; and Kraut, R. E. 2011. CrowdForge: Crowdsourcing complex work. In *Proceedings of the Symposium on User Interface Software and Technology*, UIST '11, 43–52.

Kleinberg, J. M.; and Raghavan, M. 2018. Selection problems in the presence of implicit bias. In *Innovations in Theoretical Computer Science Conference*, ITCS 2018, 33:1–33:17.

Kovera, M. B. 2019. Racial disparities in the criminal justice system: Prevalence, causes, and a search for solutions. *Journal of Social Issues*, 75(4): 1139–1164.

Lietz, P. 2010. Research into questionnaire design: A summary of the literature. *International Journal of Market Research*, 52(2): 249–272.

Miller, G. A. 1956. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2): 81–97.

Moss-Racusin, C. A.; Dovidio, J. F.; Brescoll, V. L.; Graham, M. J.; and Handelsman, J. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41): 16474–16479.

Nelsen, R. B. 2010. *An Introduction to Copulas*. Springer Publishing Company, Incorporated.

Noothigattu, R.; Shah, N.; and Procaccia, A. 2021. Loss Functions, Axioms, and Peer Review. *Journal of Artificial Intelligence Research*.

Osborne, C. 1991. Statistical calibration: A review. *International Statistical Review/Revue Internationale de Statistique*, 309–336.

Page, L.; and Page, K. 2010. Last shall be first: A field study of biases in sequential performance evaluation on the Idol series. *Journal of Economic Behavior & Organization*, 73(2): 186–198.

Pogrebtsova, E.; Luta, D.; and Hausdorf, P. A. 2020. Selection of gender-incongruent applicants: No gender bias with structured interviews. *International Journal of Selection and Assessment*, 28(1): 117–121.

Retelny, D.; Robaszkiewicz, S.; To, A.; Lasecki, W. S.; Patel, J.; Rahmati, N.; Doshi, T.; Valentine, M.; and Bernstein, M. S. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the Symposium on User Interface Software and Technology*, UIST '14, 75–85.

Saposnik, G.; Redelmeier, D.; Ruff, C. C.; and Tobler, P. N. 2016. Cognitive biases associated with medical decisions: A systematic review. *BMC Medical Informatics and Decision Making*, 16(1): 1–14.

Schmidt, F. L.; and Hunter, J. E. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2): 262.

Shah, N. B. 2022. An overview of challenges, experiments, and computational solutions in peer review. https://bit.ly/PeerReviewOverview (Abridged version published in the Communications of the ACM).

Shah, N. B.; Balakrishnan, S.; Bradley, J.; Parekh, A.; Ramchandran, K.; and Wainwright, M. J. 2016. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17: 1–47.

Shah, N. B.; Tabibian, B.; Muandet, K.; Guyon, I.; and Von Luxburg, U. 2018. Design and Analysis of the NIPS 2016 Review Process. *Journal of Machine Learning Research*, 19(1): 1913–1946.

Surowiecki, J. 2005. *The wisdom of crowds*. Anchor.

Tomkins, A.; Zhang, M.; and Heavlin, W. D. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48): 12708–12713.

Vives, M.-L.; Fernandez-Navia, T.; Teixidó, J. J.; and Serra-Burriel, M. 2021. Lenience breeds strictness: The generosity-erosion effect in hiring decisions. *Science Advances*, 7(17): eabe2045.

Wang, J.; Baharav, C.; Shah, N. B.; Woolley, A. W.; and Ravi, R. 2022. Allocation schemes in analytic evaluation: Applicant-centric holistic or attribute-centric segmented? *arXiv preprint arXiv:2209.08665*.

Wang, J.; and Shah, N. B. 2019. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '19, 864–872.