Distribution Neglect in Performance Evaluations

Eli Awtrey*†

University of Cincinnati

Nico Thornley*

INSEAD and Humu, Inc.

Jennifer E. Dannals

Dartmouth College

Christopher M. Barnes

University of Washington

Eric Luis Uhlmann†

INSEAD

* Equal author contribution

† Corresponding Authors:

Eli Awtrey, Department of Management, University of Cincinnati, PO Box 210165, Cincinnati OH, 45215-0165, Email: eli.awtrey@uc.edu

Eric Luis Uhlmann, Organisational Behaviour Area, INSEAD, 1 Ayer Rajah Avenue 138676, Singapore, Email: eric.luis.uhlmann@gmail.com

CRediT authorship contribution statement

Eli Awtrey: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Nico Thornley**: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Jennifer E. Dannals:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Christopher M. Barnes:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Eric Luis Uhlmann:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

Funding Acknowledgments

Eric Luis Uhlmann is grateful for an R&D grant from INSEAD in support of this research.

Abstract

Five empirical studies, including both laboratory experiments and an archival investigation, provide evidence that decision makers often fail to consider variability and skew when making judgments about performance. We term this *distribution neglect*. Participants' spontaneous explanations for group differences in elite achievement overwhelmingly invoked mean differences rather than group differences in variability, even when the complete distribution and summary statistics were provided (Study 1). A longitudinal examination indicates that NBA teams overweight average performance and underweight consistency of performance when deciding players' contracts (Study 2), providing evidence that neglecting variance information leads to suboptimal judgments. In a manufacturing scenario involving monitoring assembly line workers, participants were more accurate at identifying top (high mean) performers than consistent (low variability) performers (Study 3). In a hiring simulation, decision makers were more likely to factor in variance when performance data was presented visually as a histogram (Study 4). Finally, participants' spontaneous explanations for others' self-assessments of ability assumed egocentric bias, when a skewed performance distribution was also a plausible contributor (Study 5). Individual differences (need for cognition) and task differences (such as style of information display) were associated with increased distribution-based reasoning in multiple studies, suggesting potential boundary conditions for further investigation. Organizational implications, and additional potential remedies for distribution neglect, are discussed.

Keywords: Performance evaluations; selection decisions; compensation; intuitive statistics; decision heuristics; distribution; dispersion; variance; skew

Statistical intuitions are often systematically biased (Ayton & Fischer, 2004; Clotfelter & Cook, 1993; Croson & Sundali, 2005; Kahneman & Tversky, 1979; Peterson & Beach, 1967; Sunstein, 2002; Tversky & Kahneman, 1974). People prefer to conserve cognitive resources (Fiske & Taylor, 1991). Thus, instead of complicated mathematical calculations, individuals rely on heuristic approximations that are accurate in many but not all instances (Gigerenzer, 2008; Kahneman, 2011). A number of these biases reflect, at least in part, intuitive misunderstandings about variance. Failure to account for regression to the mean suggests that people often account for mean differences but not natural and expected variations in data, a mistake made by trained researchers (Halliday, Thomas, Siu, & Allison, 2018) and experienced organizational decision makers (Paola & Scoppa, 2012). Similarly, even psychology students and trained researchers can misunderstand variance in confidence intervals (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). In addition, people neglect variation in social groups, particularly when stereotyping outgroups as homogeneous (Quattrone & Jones, 1980) and may attend to the valence of discrepant scientific findings but not their extremity (Fisher & Keil, 2018). Variance and asymmetry can provide critical information about a population, yet there are many examples of people discounting these statistics even when they are directly relevant. How and when do people use distributional information in data-based decision-making?

Improving data-based decision-making is a critical societal goal, particularly in the realm of performance evaluation, where decisions are often biased. Managers do not sufficiently discriminate between areas upon which an employee is judged and instead tend to give a global rating to the employee and apply it to each evaluation, demonstrating a halo effect bias (Holzbach, 1978). This bias is a dominant predictor of variance of performance ratings (Klimoski & London, 1974; Prien & Liske, 1962). Managers also often relying on the typical or

modal employee performance (Barnes & Morgeson, 2007; Dubois, Sackett, Zedeck, & Fogli, 1993; Lecerf, Ghisletta, & Jouffray, 2004; Sackett, Zedeck, & Fogli, 1988) and may be biased by performance trends, overweighing recent performances and performance trajectories when it comes to deciding compensation (Barnes, Reb, & Ang, 2012; Ferris, Reb, Lian, Sim, & Ang, 2018).

Distribution Neglect

We propose a general cognitive tendency that can distort performance evaluations: *distribution neglect*, defined as the tendency to misestimate and underutilize distribution information. We argue that this arises from an overreliance on mean information. Underutilizing variance and skew and instead relying more on statistical means can prove an efficient heuristic, but we propose it can lead to systematically suboptimal judgments. We do not suggest that distribution information is more important than mean information or that it is entirely ignored, but rather that it is comparatively neglected and thus may lead to systematic failures in data-based decisions.

A probability distribution is a mathematical function that describes the probability of future events based on a given data generating process (Parzen, 1962). In a normal probability distribution, prediction relies on the mean because the probability of a future event from the same data generating process is greatest at the mean. This is not only because the mean outcome has the highest probability of occurrence for normal distributions, but also because mean-based predictions will often have relatively low expected error. Therefore, means are generally good predictors of future events.

Normal distributions are so common that many statistical techniques assume normality even when the shape of the distribution is unknown (Casella & Berger, 2001). Nonetheless, even

in highly skewed distributions (e.g., power law distributions), mean-based predictions naturally balance (a) making a prediction that represents the highest probability outcome with (b) minimizing the expected error in the prediction. For example, if a loaded die were to land on "6" for 70% of rolls but was otherwise uniformly distributed across the other 5 numbers, then the average expected outcome would be 5.1. Therefore, a prediction of 5.1 (or 5 if an integer is necessary) for future rolls would both keep predicted outcomes close to actual outcomes and also minimize error distance. Like many heuristics, a mean heuristic may be a poor predictor in less common circumstances. For example, the mean would not be as useful for predicting future outcomes of certain multimodal distributions, though they still outperform medians in a bimodal distribution (Mosteller & Tukey, 1977). However, bimodal distributions are uncommon in nature, with unimodal distributions (e.g., normal or power law distributions) more common (Casella & Berger, 2001; Gabaix, 2016), again reinforcing the utility of mean-based predictions.

Additionally, the utility of mean predictions is robust to sample size. With a single observation, the mean (i.e., the single observation) is still the best estimate of the population mean because, without more data, one has no additional information about how much the data generating process might deviate or in which direction. As the sample size increases, the sample mean adjusts to become a more reliable indicator of the underlying population mean and therefore generally becomes a more useful predictor of future events.

Using the mean outcome for prediction thus improves accuracy, reduces error, and avoids complex considerations of variance information. This does not imply that means are the ideal method for estimating future events produced by the same data generating process, only that the benefits of using mean-based predictions (e.g., accuracy, reduced complexity) make them relatively efficient and effective. Consequently, a mean heuristic could be a reasonable

approximation of future events across a variety of statistical situations. People may also prefer means because they are computationally easy to estimate, can be adjusted incrementally after new observations, and reduce the need to have a complete memory of previous events thereby reducing demands on memory. However, these features of means also suggest that individuals may become overly reliant on means and thereby underutilize other distributional information in judgment, similar to many other observed cognitive tendencies that reduce cognitive demand but sacrifice some accuracy (Gigerenzer, 2008; Kahneman, 2011). Because of a focus on means during intuitive processing of performance information, other information about the distribution may tend to be neglected. Under cognitive demands, where heuristics are most likely to be in use, this means that individuals would focus more on the mean, and less on other statistics (e.g., standard deviation, skew). Though potentially effective in some cases for the reasons outlined above, we suggest it this tendency may also lead to predictable errors.

More deliberate analyses of distribution information will remain challenging for the same reason that an intuitive understanding of distributions may be underdeveloped in humans: computational difficulty. We therefore expect that even when decision makers are directly told to attend to and calculate distributions, they will do a better job estimating means than they do estimating variances. People cannot be equally good at every computational task, and for human beings, calculating means may be less complex and more intuitive than calculating variances.

At the same time, there are other computationally simple methods for understanding distribution information. Providing a full distribution of performance scores and/or summary statistics like standard deviation should prompt distribution-based reasoning, relative to observing subsets of the distribution. Further, visualizing a distribution (for example with a histogram) could make it easier to imagine how the variance could change or fluctuate. This task

becomes easier by engaging more robust visual processing resources that do not require tedious and complex mathematical computations, such as calculating variance or standard deviation (Resnick, 1987; Scribner, 1984; Wheatley, 1991). Moreover, a visual representation of a distribution provides information more intuitively by representing the entire distribution in a single snapshot, displaying complex information in an image that can be digested almost instantaneously (Thorpe, Fize, & Marlot, 1996).

To summarize, we hypothesize the following:

Hypothesis 1: Mean information will be used more than distribution information (variance and skew) when attempting to explain differences in performance.

Hypothesis 2: Distribution information is underutilized relative to the normative benchmark of what would maximize the accuracy of judgments.

Hypothesis 3: Individuals will estimate distributions less accurately than averages.

Hypothesis 4: Increasing the completeness and salience of distribution information will increase its use in explaining differences in performance.

Beyond these four hypotheses, our empirical studies also collectively propose and test a taxonomy for reasoning about distributions, with some forms of variability hypothesized to be intuitively easier to grasp than others. We suggest that people first tend to intuitively explain the world in terms of means rather than in terms of variance. Although variance is far from completely ignored, it is systematically underweighted relative to mean information and also relative to what would maximize judgmental accuracy. When people are motivated to think about the full distribution, we further anticipate they tend to implicitly assume data is normally distributed rather than skewed. Individuals acting as intuitive statisticians may make the same assumption that many statistical tests do (Casella & Berger, 2001) when reasoning in everyday

life. This assumption would be adaptive to the extent than many distributions are in fact normal but would also yield predictable errors when individuals fail to take potential skew into account. Mean, standard deviation, and skew are each in turn less intuitive, and therefore progressively less likely to be factored into everyday decisions. This pattern should be moderated by expertise in the domain, as well as statistical training more generally, both of which make it easier to process progressively more complex and less intuitive forms of dispersion. Contrarily, heuristic processing, whether situational or chronic, should lead decision makers to "devolve" towards simpler ways of reasoning about variance, and in some cases rely solely on means as a heuristic.

Empirical Overview

Across five studies, we provide converging evidence of distribution neglect using both experimental and archival methods (Barnes, Dang, Leavitt, Guarana, & Uhlmann, 2018) and with both non-expert (Studies 1, 3, 4, and 5) and expert (Study 2) decision-makers to examine whether the effect disappears under conditions of accountability or with greater experience in the domain. Study 1 tests if participants fail to spontaneously consider variance-related causes for group differences in performance, and if providing more complete information can help address this. Study 2 analyzes twenty-five seasons of data from the National Basketball Association (NBA), examining how professional basketball teams weight consistency of performance relative to mean performance when deciding player compensation. Study 3 prompts participants to consider means and variances in a workplace performance evaluation scenario, testing whether individuals are less accurate when identifying differences in variances than differences in means. Study 4 explores whether visual presentation of distributions can attenuate variance

neglect. Study 5 tests if participants neglect to think of potential skew-related contributors to assessments of performance, and whether visual presentation is an effective intervention.

These studies represent initial investigations of distribution neglect and our proposed taxonomy. We discuss potential boundary conditions, remaining questions, and future directions in the General Discussion. Increasing the informational value of these experiments, Studies 1, 2, 3, and 5 were pre-registered prior to data collection, and all five studies feature open data and materials. Information on pre-registration plans, survey instruments, anonymized data, analysis code, and pilot studies can be found in the Online Supplements and at https://osf.io/923n6/.

Study 1: Explanations for Group Differences in Performance

We examined if people generate mean-type responses earlier and more frequently than variance-type responses as explanations for group differences in competitive performance. We also test if quantity of information presented affects the use of variance reasoning by randomly assigning participants to conditions that vary the quantity and type of information presented. We also examined moderation by individual-differences including the Need for Cognition (NFC) or chronic tendency to process information deliberatively (Cacioppo & Petty, 1982), and the number of statistics courses participants had previously completed.

Methods

Participants. We recruited 600 respondents via Prolific.co, requiring residence in the United States and fluency in English. Participants received $1.50 for the 10-minute survey. After cleaning the data for inattentive behaviors as outlined in the pre-registration (e.g., failing an attention check), the final sample size was N = 553 (92.2%). The median age of participants was 27 years, and 53% of the sample self-identified as female.

Procedure and conditions. Participants were presented with a scenario in which a planet with alien life had been discovered with many species of different shapes, sizes, and colors (see Supplement 1) and an experiment had been conducted in which "100 individual aliens were picked at random from the total population of 2 different species," and the selected individuals competed in a 100-yard dash. Participants were then randomly assigned to one of five conditions that represented different display forms of the racing data (see Supplement 1 for a depiction of these displays.) All conditions used the same underlying dataset but presented different parts of it. Both species had the same group mean time (15 seconds) but differed on the standard deviation of race times, such that the species with more top finishers had a higher standard deviation (3.3 vs. 2.4).

Condition 1: Top 10 places. Participants were shown the top ten finishers (#1-10) out of 100. (A pilot version of this study featuring only Condition 1 is reported in Supplement 3.)

Condition 2: Top 10 and bottom 10 places. Participants were shown the top ten finishers and the bottom ten (#91-100) finishers.

Condition 3: Full distribution. All 100 finishers were presented to participants.

Condition 4: Full distribution with summary statistics (mean then SD). In addition to race times, participants also received summary statistics for each species, mean and standard deviation, with mean presented first.

Condition 5: Full distribution with summary statistics (SD then mean). In addition to race times, participants also received summary statistics for each species, mean and standard deviation, with standard deviation first.

Measures

Explanation type and order. Participants were asked to list reasons that would be likely to produce the results that were shown, in the order in which the explanations came to mind. Independent raters blind to condition and hypotheses used a pre-registered coding scheme to categorize the open-ended responses into reasoning based on (1) differences in group means, (2) differences in group variance, or (3) differences in population size. Codes for (4) vague, (5) off-topic and (6) multiple were also included. The last code was used when a statement invoked multiple types of explanation at once. The ordered position of the explanation was also coded, in order to examine the types of explanation that were most likely to come to mind first.

Mathematical and statistical proficiency. Participants were asked to self-rate their mathematical and statistical proficiency on a scale from 1-10, with 1 being "extremely low" and 10 being "extremely high."

Statistical courses. Participants were asked how many courses on statistics they had taken in their life.

Need for cognition. Participants completed the brief 18-item need for cognition scale (Cacioppo, Petty & Kao, 1984). Alpha for this scale was .92.

##

INSERT TABLES 1 AND 2 ABOUT HERE

##

Results

Primary Analyses. Count and percentage data for the different categories of reasoning are provided in Table 1. Per the preregistered analysis plan, we used an exact binomial test to evaluate the degree to which distributional reasoning was used relative to mean reasoning.

When evaluated across all conditions, mean reasoning was used significantly more than all other categories of reasoning (0.644, CI 95%[0.620,0.667], $p < .001$) and was also used significantly more than variance reasoning when compared pairwise (0.943, CI 95%[0.928,0.956], $p < .001$). These analyses were also evaluated within each condition with a similar pattern of results (see Table 2). This supports Hypothesis 1, which predicted that distribution information is used less often than mean information when attempting to explain group differences in performance.

##

INSERT TABLE 3 ABOUT HERE

##

To evaluate the effect of information condition on distribution neglect, we pre-registered an ordinary least squares regression model in which the category of the first reason provided (0 = not variance reasoning, 1 = variance reasoning) was regressed on condition with the *Top 10 places* condition being the base case. (While a logistic regression model is traditionally used with binary outcome variables, an OLS regression is suitable for these kinds of outcomes in experimental data; Gomila, 2020). Results with a logistic regression model yield the same pattern of results and are available in Supplement 2. The model evaluates whether providing increased information to participants is associated with changes in variance reasoning and therefore distribution neglect. Robust standard errors with clustering on the respondent were used to evaluate statistical significance. The results of this model are shown in Table 3. Relative to the base case of *Top 10 places*, variance reasoning was more often used when participants were further provided the bottom of the race results ($b = 0.06$, SE = 0.02, $p < .001$), all of the race results ($b = 0.03$, SE = 0.01, $p = .001$), all of the race results plus the mean then standard deviation ($b = 0.06$, SE = 0.02, $p < .001$), or all of the race results plus the standard deviation

then the mean (b = 0.04, SE = 0.01, p = .002). Each information condition reduces distribution neglect relative to only providing the top 10 places, which provides support for Hypothesis 4 that greater information completeness and salience increases reliance on distribution-based reasoning.

Secondary analyses. Also as listed in the preregistered analysis plan, as a secondary analysis we investigated the role that various individual differences played in predicting distribution neglect. We tested separate OLS models to evaluate the degree to which each individual difference of interest was associated with the coded categorization of the first reason provided by respondents. When considering all responses, neither need for cognition (b = 0.01, SE = 0.01, p = .089), self-rated math reasoning (b = 0.01, SE = 0.00, p = .080), self-rated statistical reasoning (b = 0.00, SE = 0.00, p = .519), nor number of statistical courses taken (b = 0.00, SE = 0.00, p = .370) was significantly associated with the use of variance reasoning. However, when considering only the first response, need for cognition was positively and significantly associated with the use of variance reasoning (b = 0.03, SE = 0.01, p = .011), but self-rated math reasoning (b = 0.01, SE = 0.01, p = .156), statistical reasoning (b = 0.0, SE = 0.01, p = .517), and number of statistical courses (b = 0.00, SE = 0.00, p = .226) were not significantly associated with the use of variance reasoning.

We also expected in our pre-registration that the *Top 10 and bottom 10 places* condition would generate more population reasoning than other conditions. One intuitive explanation for numerous top and bottom performers, not knowing the middle of the distribution, is that the group in question is simply more numerous. Population reasoning assumes that the reason for overrepresentation of one species in the race results is due to a different quantity of racers from the two species, which represents a different form of distribution neglect than is represented by a reliance on mean reasoning. Using an OLS regression model with robust standard errors

clustered on the respondent, we tested each condition against the *Top 10 and bottom 10 places* condition as a base case. As predicted, *Top 10 places* (b = -0.12, SE = 0.03, $p < .001$), *Full distribution* (b = -0.14, SE = 0.02, $p < .001$), *Summary statistics with mean then SD* (b = -0.15, SE = 0.02, $p < .001$), and *Summary statistics with SD then mean* (b = -0.14, SE = 0.02, $p < .001$) were all significantly less likely to invoke population differences than the base case.

We also examined potential moderating relationships between individual differences and our experimental conditions. Using an OLS regression (again with standard errors clustered by individual), only a few of the interaction terms were statistically significant at $p < .05$, providing little consistent evidence that need for cognition, self-rated math reasoning, self-rated statistical reasoning, or past training in statistics moderates the negative relationship between information availability and distribution neglect. However, NFC did predict the use of variance reasoning in the first response given, regardless of condition. This suggests that individuals high in need for cognition are more likely to take into account variance or standard deviation, regardless of whether a high or low quantity of information is available.

Finally, we pre-registered an analysis of a linear relationship between increasing information condition (as ordered above in the Methods section) and increased distribution reasoning (or decreased distribution neglect). An OLS regression (robust standard errors clustered on participant) with variance reasoning as the dependent variable and ordered condition (ranging from 1-5) as the independent variable showed a positive and significant relationship (b = 0.01, SE = 0.00, p = .006). This adds additional support for Hypothesis 4, in that distribution-based reasoning increased with information availability and salience.

Discussion

In Study 1, people typically failed to spontaneously generate variance-related explanations for performance differences that could be plausibly explained by variance (e.g., higher standard deviation in one population than the other, such that they have more very fast and very slow members), offering evidence for Hypothesis 1. The observed effect is not necessarily a bias— indeed, it is arguably quite reasonable to infer that a group overrepresented among winners is faster, but it is notable that this is a default. As expected, distribution-based reasoning increased when the entire set of scores was visible and standard deviations were explicitly provided, yet never approached the frequency of means-based reasoning. Even when shown the full set of scores and told the standard deviation prior to the mean, 190 explanations for the race results invoked mean differences and only 13 invoked variance differences. The most common explanation used means rather than variance under not only information-poor circumstances (Condition 1) but also information-rich circumstances (Condition 5), an important first step in establishing that mean-thinking dominates considerations of variance or skew. However, Study 1 does not demonstrate underutilization of distribution information relative to optimal use in the real world, which we will examine in Study 2.

The scenario in Study 1 mirrors real-life domains such as athletics and intellectual achievement, in which top performers are most salient, and demographic variables are likewise salient in that performers are grouped psychologically by race or gender (Greenhaus & Parasuraman, 1993). In such situations, observing just the top performers can lead people to attribute differences in variance to group differences in mean. As a result, when there are differences in variance across demographic groups, distribution neglect could contribute to the formation of stereotypes about the typical or average members of social groups. If so, uncovering

factors that reduce distribution neglect may be useful for combating this kind of stereotyping. We return to the topic of interventions to encourage distribution-based reasoning in later studies as well as the General Discussion.

Study 2: Underutilization of Variability Information in Setting NBA Salaries

In this study, we tested Hypothesis 2 in a context in which decision makers would be highly familiar with performance statistics and also accountable for their judgments—compensation decisions made by general managers of professional sports organizations. We examined how professional basketball teams in the National Basketball Association (NBA) assess the value of players. Players with higher average performances help their teams to win more games and are therefore better remunerated by the organization (Barnes & Morgeson, 2007; Reb & Cropanzano, 2007; Zhou & Martocchio, 2001). At the same time, in an interdependent task that requires multiple individuals to complete a string of successful actions, performing consistently is important for a team's success (Barnes et al., 2012). Therefore, players that have a lower standard deviation in performance (i.e., greater consistency) should also earn a higher salary (Barnes & Morgeson, 2007). Consequently, in this study, we test if NBA teams display distribution neglect by evaluating if (a) variability of performance receives less weight in determining compensation than it does in determining team performance and (b) performance mean receives more weight in determining compensation than it does in determining team performance. If so, this would highlight the consequences of neglecting dispersion information in evaluations, by suggesting that NBA teams suboptimally underuse distribution information.

Methods

Design and sample. Game and salary data from 25 NBA seasons were collected from the 1995-96 season through the 2019-2020 season (a pilot version of this study including with a smaller sample of seasons is reported in Supplement 4). Data were evaluated at both a player level and at a team level to allow for models that predict player salaries and team wins respectively. For the individual level dataset, we began with a sample that included all 1478 players drafted in the first two rounds of the NBA draft (in which new players to the league are selected by each team) from the summers of 1995-2019. However, for a variety of reasons, many drafted players do not end up playing in the NBA. Further, many of the players that do initially play in the league do not receive a second contract after their initial rookie contract. This is an important distinction, because we wanted to focus on players whose contracts were based on the evaluation of prior performance records. Thus, our sample included only players that played in the NBA long enough to receive a second contract, reducing our sample to 727 players in 43,159 player-game pairings. For the team level dataset, our interest was in evaluating the true effect of player performance means and variability on team outcomes (wins). Thus, we included all games and all players over the 25-year observation window (1995-96 season through 2019-2020 season) for a total of 741 team-season pairings in 29,417 games.

Measures: Individual level.

Player salary. Player salary was evaluated as the salary earned by the player in the first full season after signing the new contract. Signing bonuses were not included in the salary figures used in our analysis because they are uncommon in the NBA. Only 16 out of the top 2000 contracts over our timespan (less than 1%) included signing bonuses, and only 7 of these were worth more than 10% of the contract value (Sportrac.com, 2021).

Player performance. Player performance was evaluated as the season of game statistics he accumulated prior to the contract signing. This "Game Score" metric is widely used for quantifying the quality of a player's game performance and generates a single performance score by weighting a number of recorded in-game actions based on their relative value for the team (Hollinger, 2003, 2005). Specifically, the Game Score is computed with the following equation:

*Game Score = Points Scored + (0.4*Field Goals) - (0.7*Field Goal Attempts) - (0.4*(Free Throw Attempts-Free Throws)) + (0.7*Offensive Rebounds) + (0.3*Defensive Rebounds) + Steals + (0.7*Assists) + (0.7*Blocks) - (0.4*Personal Fouls) - Turnovers*

We computed this for each player for each game, and then computed season-level statistics (mean and standard deviation) for each player (mean number of games = 59.37).

Control variables. We controlled for experience by including variables for player age and tenure (in number of NBA games), which have been shown to be significantly related to player salary (Barnes & Morgeson, 2007). To account for categorical differences in salary, we controlled for position (dummy variables for Forward or Center with Guard as a base case) and if the player was a free agent (changed teams in the previous year) (Barnes et al., 2012). To mitigate recency effects, we controlled for the linear trend of the focal player's performance over the course of the season, as this has been shown to predict evaluations of employee performance in prior work (Barnes et al., 2012; Reb & Cropanzano, 2007; Reb & Greguras, 2010). Finally, we controlled for prior player salary and included fixed effects (with dummy variables) for season.

Measures: Team Level.

Team performance (wins). This was operationalized as the number of wins a team had in each season.

Players' performance mean and standard deviation (team level). Again, Game Score is used as the metric for player performance, but for team level models this individual player performance is aggregated to the team level. Since the amount of playing time given to different players varies widely, the impact of any one player's performance on team outcomes will also vary. Thus, in these models we have used minute-weighted averages of player performance means and standard deviations as our method of aggregation to account for this differential effect.

Controls. As in the individual level models, we include dummy variables for season to control for annual differences (e.g., different distributions of win-loss records, shortened seasons).

Results

Tables 4 and 5 display the means, standard deviations, and zero-order correlations for variables in the individual and team models respectively, and Tables 6 and 7 show the regression results for these two models.

##

INSERT TABLES 4 AND 5 ABOUT HERE

##

In the context of this study, player salaries are indicators of predicted future player performance—individuals that are expected to have better future performance should be more highly compensated, controlling for factors such as position scarcity, market conditions by year, and other potentially confounding variables. To this point, we found that player salaries were significantly related to individual player performance means, both in a zero-order correlation (r =

.53, $p < .01$) and in regression models (Table 7, Model 3: $b = 321456.31$, $SE = 56470$, $p < .001$, $\beta = .34$).

##

INSERT TABLES 6 AND 7 ABOUT HERE

##

In order to establish an appropriate utilization of distributional information, we first evaluated the effects of aggregated players' performance (both means and standard deviations) on teams wins. These models (Table 6) indicated that team wins were predicted by players' performance standard deviation (Model 3: $b = -14.12$, $SE = 1.12$, $p < .001$, $\beta = -.34$) in addition to players' performance mean (Model 3: $b = 10.13$, $SE = 0.36$, $p < .001$, $\beta = .76$). It would follow that player salaries would also take both factors into account. However, when predicting player salary (Table 7), individual player performance mean was significantly related to the outcome (Model 3: $b = 321456.31$, $SE = 56470$, $p < .001$) but individual player performance standard deviation was not (Model 3: $b = 34386.73$, $SE = 148165$, $p = .817$) when both are included in the model. Further, a Wald test comparing individual player performance mean and individual player performance standard deviation in the player salary model was non-significant ($\chi^2 = 2.19$, $p = .139$), suggesting that dispersion information did not add any explanatory power when predicting player salaries. Conversely, the Wald test comparing team-level players' performance mean and team-level players' performance standard deviation in the team wins model was significant ($\chi^2 = 375$, $p < .001$), indicating that both team-level players' performance mean and team-level players' performance standard deviation were substantive in predicting team wins. This same pattern is seen in changes in variance explained when adding performance standard deviation to the regression models, which were significant in the team wins model ($\Delta R^2 = .094$, $p < .001$) but

not in the player salary model (ΔR^2 = .000, p = .817). Finally, the standardized regression coefficient for team-level players' performance standard deviation in the team wins model (Model 3: β = .34) was much larger than the corresponding individual coefficient in the player salary model (Model 3: β = .01). Put together, these findings provide support for Hypothesis 2, that distribution information is underutilized relative to what would optimize judgmental accuracy.

Discussion

These results underscore how important it is to consider both mean and variance information when making judgments, in that both team member mean performance and consistency significantly predict team success. And yet, NBA teams underweight the importance of variability in performance when estimating the value that players contribute to the team. This provides evidence of distribution neglect in a naturalistic environment among expert decision makers under conditions of high accountability. Variance information is clearly valuable, and the professional sports teams in this sample do not appear to use it as much as they should in light of its impact on team performance. Managers neglect to fully consider variance when setting player compensation, potentially hurting their roster building and team performance. This supports Hypothesis 2, that people systematically underutilize distribution information in their performance judgments, leading to suboptimal results.

Variability is especially relevant in team contexts such as this in which people rely on each other. High variability in team member performance harms predictability, coordination and collaboration. This makes basketball, an interdependent team sport (Swaab et al., 2014), precisely the sort of context in which decision makers should be most closely attuned to variance in performance, providing a conservative test of the hypothesis. If any sports managers should

appreciate the importance of variability, it should be basketball managers. However, this interdependency complicates our assessment of individual performance and its inconsistencies. For example, interpreting the relationship between performance inconsistency in year 1 and salary offered in year 2 is admittedly tricky in basketball. Future research might examine distribution neglect in comparatively independent sports such as baseball, which provide cleaner individual-level measures of performance, and might also directly compare interdependent and independent sports (Swaab et al., 2014).

It is also worth considering how salary cap constraints might affect our findings. In NBA basketball all teams follow the same salary cap rules each year, such that this system is a constant across teams. Further, we included fixed effects for year, so our findings should be robust to variance due to annual changes in salary cap rules. However, it is worth noting that salary caps put constraints on how managers can allocate compensation, making it even more important to allocate compensation to maximize return on investment. Suboptimal compensation creates high opportunity costs. Notably, this again makes basketball a conservative test of distribution neglect, because those managers should be more motivated than those under lessened financial constraints to pick up on the value of (low) variability.

Two alternative explanations merit additional consideration. First, it is possible managers hope that a volatile player will evolve into a consistent superstar and therefore temporarily discount the variability in performance. In other words, managers may bet that a player with a high standard deviation in his game-to-game performance will ultimately be able to reduce this variance and offer a top contract. Some indirect support for this idea comes from Reeder and Brewer (1979), who show that top performances are seen as more diagnostic of ability than low performances, since a top talent will perform poorly sometimes whereas someone without talent

will never perform well (see the General Discussion for potential links between schematic models of dispositional attribution and distribution neglect.) Second, high-dispersion players may add unobserved value if a peak or bravura performance could sell merchandise more than a temporary slump in performance hurts along this same dimension. Notably, this would still reflect distribution neglect on the part of fans, albeit not team managers. The relevant data on merchandising revenues per NBA player are not publicly available, so we must leave parsing specifically who is exhibiting distribution neglect (i.e., managers, fans, or perhaps both) to future research. Our experimental studies are less subject to these counter-explanations and thus, the strengths of our experimental and archival studies complement each other and compensate for the weaknesses of each respective methodology (Barnes et al., 2018). Future research should use archival datasets to explore the role of performance variability in predicting performance outcomes as well as selection, promotion, and compensation levels across further industries.

Study 3: Accuracy in Identifying Top Performers vs. Consistent Performers

Study 1 found that people were less likely to spontaneously identify variance-related causes of performance outcomes than mean-related causes when unprompted, and Study 2 documented suboptimal use of performance variance information relative to what would maximize outcomes in a real-world performance setting. In Study 3 we tested for distribution neglect in a workplace performance scenario in which we specifically prompted participants to look for dispersion differences. We ask participants to compare the real-time quality ratings of two assembly lines to supply employees with feedback on the lines with the higher mean ratings or the greater consistency of ratings. This enables us to test if participants are also less accurate in correctly identifying variance information when specifically directed to look for it. Participants made performance judgments under time constraints, similar to how some

manufacturing settings require quick and ongoing assessments of work output by supervisors (SIOP, 2014; Miller, 2019). As in Study 1, we assessed individual differences such as need for cognition as well as number of courses in statistics previously completed. A pilot version of this study is reported in Supplement 7.

Methods

Participants. We recruited 600 respondents via Prolific.co, requiring residence in the United States. Participants received $1.15 for the 7-minute survey and were eligible for up to $0.50 in bonus payments. One participant quit the survey early, leaving an initial pool of 599 participants. After cleaning the data for inattentive behaviors as outlined in the pre-registration (e.g., failing an attention check), the final sample size was N = 545 (90.8% of those recruited). The median age of participants was 32 years, and 51% of the sample self-identified as female.

Procedure and conditions. Participants were told that they were in the role of a manufacturing supervisor in an electronics factory. Throughout the day, they quickly peek at the quality ratings on each assembly line to provide workers with real-time feedback on their performance. Participants were then shown ten sets of "quality ratings", which were paired data distributions randomly chosen from 200 possible pairs (see Supplement 5). Participants were to then choose within each set the distribution that either reflected the "higher overall average quality rating" if the participant was assigned to the mean condition or the "more consistent quality rating" if participant was assigned to the variance condition. Note that participants were asked to identify the more "consistent" output, not the more "reliable" output. The latter adjective has more positive connotations and could be taken to mean a generally good output, i.e., both high mean and low variance.

Participants made judgements about ten pairs of assembly lines and were paid a $0.05 bonus per correct judgement. For each judgment, participants could view the paired data for a total of 10 seconds. If the participant had not selected a decision by the end of that time period, the survey then automatically advanced and prompted participants for their choice. All participants completed one trial round so that they were familiar with the procedure.

This experimental paradigm drew heavily on past approaches from Reb and colleagues (Reb & Cropanzano, 2007; Reb & Greguras, 2010), which rely on relative judgments between targets. The reason for doing so is that otherwise participants will not have a sound basis from which to judge whether an absolute value is high or low. Making the comparisons relative simplifies the decision task for the participants, in that it removes that particular source of ambiguity. In other words, we sought to minimize participant confusion regarding what is a "high" number on an absolute scale.

Measures

Decision accuracy. For each judgement, participants were given a score of 1 if they chose the correct assembly line (the one with the higher mean performance in the mean condition or the lower standard deviation in the variance condition) and a score of 0 if they chose the incorrect assembly line.

Decision simplicity. For each possible set of two performance distributions, we calculated how objectively simple the task of assessing mean performance versus consistent performance would be in order to generate a normative benchmark. To be clear, we sought to put mean and variance comparisons on equal footing mathematically. However, this does not mean that human participants would not still find variance harder to calculate than means, only that we have controlled for how difficult a purely rational artificial intelligence would find these tasks.

To the extent human participants find variance more intuitively difficult to calculate than means, even when the two tasks would be equally difficult for an artificial intelligence, our theoretical Hypothesis 3 regarding distribution neglect is supported.

Because mean differences are normally distributed, while variance differences are F-distributed, absolute difference in means are not directly equally difficult to judge as compared to absolute differences in variances. Calculating objective decision simplicity thus allows us to better directly compare the task of judging average performance versus consistent performance. To do this we ran two statistical functions per pair of distributions. The first ("pnorm" in R) tells us the probability that the difference in means between the higher distribution and the lower distribution is greater than or equal to a 0.1 difference in performance. The second ("pf" in R) tells us the probability that the ratio of means is not equal to 1, which is to say, the likelihood that one distribution has significantly greater variance than the other. One can conceptualize these two tests as asking: "If a computer with infinite capacity of computation were to encounter the problem we set for participants, how different would it find these two distributions with respect to differentiating their means and differentiating their variances?" Again, this does not control for any psychological tendencies a participant might have in computing mean or variance, because the computer would have no such biases, but does control for the strict computational differences in calculating mean and variance because we can equalize the probability that one mean is greater than the other, to the probability that one variance is greater than the other. Because higher probabilities equate to easier choices, this allowed us to control for the computational difficulty of the two tasks. A decision simplicity score of 0.99 thus indicated a very easy choice, while a decision simplicity score of 0.51 would indicate a very difficult choice. Values of the distributions used for participants varied from 0.58 to 0.99.

Statistical courses, need for cognition, mathematical and statistical proficiency. Measured as in Study 1.

##

INSERT TABLE 8 ABOUT HERE

##

Results

Primary analyses. As predicted, participants in the variance condition had a lower accuracy rate (68.0%) than those in the mean condition (72.1%). Per the preregistered analysis plan, we used a logistic regression model to test the statistical significance of this difference. (Ordinary-least squares models were also evaluated, and they produced almost identical results, as described in Supplement 6.) Robust standard errors with clustering on the respondent were used to compensate for non-independence of the data. These results of this analysis are presented in Table 8. The coefficient for the variance condition (base case being the mean condition) is negative and significant at the $p < .05$ level with decision simplicity included as a control (Model 3: $b = -0.30$, SE = 0.07, $p < .001$) or without it (Model 2: $b = -0.20$, SE = 0.07, $p = .007$). Together, these findings support Hypothesis 3, which predicted that individuals are less accurate at estimating variance than they are at estimating averages.

##

INSERT TABLE 9 ABOUT HERE

##

Secondary analyses. Also as listed in our preregistered analysis plan, we investigated potential individual differences in distribution neglect. Using the same logistical regression model used in Models 1-3, we tested separate models to evaluate the degree to which each

individual difference of interest was associated with improved decision accuracy. Controlling for decision simplicity and condition, neither self-rated mathematical reasoning (b = -0.02, SE = 0.02, p = .395) nor self-rated statistical reasoning (b = -0.03, SE = 0.02, p = .127) were significantly related to decision accuracy and thus were not evaluated further. Having previously taken more courses in statistics was unexpectedly associated with *less* decision accuracy (b = -0.05, SE = 0.02, p = .015), a pattern that did not emerge in any other study and is therefore not interpreted here. Need for cognition (see models in Table 9) displayed a marginally significant relationship with decision accuracy (Model 4: b = 0.10, SE = 0.05, p = .056) and a statistically significant interaction effect such that higher levels of need for cognition weakens the negative relationship between condition and accuracy (Model 5: b = 0.21, SE = 0.10, p = .042). Thus, higher NFC is associated with less of an accuracy loss between the mean and variance conditions. In other words, high-NFC individuals do better at reasoning about variance, but not means, relative to low-NFC individuals. See Supplement 6 for data visualizations of the relationship between NFC and judgmental accuracy in this study.

Discussion

Even though machine computational difficulty was equivalent across the variance and mean tasks, Study 3 finds that human participants fail to identify variance differences as accurately as mean differences in a workplace performance evaluation scenario. This result was found despite explicitly drawing attention to dispersion and incentivizing correct responses, suggesting that even motivated participants struggle to assess variance information. This further suggests that distribution neglect could undermine the ability of decision makers in organizations to rationally and fairly assess performance when consistency in performance is a relevant quality in evaluations.

Study 4: Reducing Variance Neglect with Visual Representations

In Studies 1-3, we observe distribution neglect in both controlled experiments and real-world settings. Study 1 found that presenting more complete information (i.e., the entire distribution and/or summary statistics) encouraged variance-based explanations for group differences in performance. Jung and Kahn (2014) report evidence that animated pictographs are more effective at communicating variance than the boxplots from Medicare websites in the United States. We therefore expected that people who view data as a histogram should be able to reason more intuitively about variance and consequently utilize such information more than when the same data is presented in table form.

Methods

In a paradigm similar to Study 3, participants acted as supervisors and determined which of two employees performed better.

Participants. We recruited two hundred and ninety-seven participants on Amazon's Mechanical Turk for a 10-minute study paying $0.75. The sample was 51.2% female, and 0.69% self-categorized as other than female or male. Participants had a median age of 34 years. We screened out participants who had completed fewer than 100 HITs at less than a 95% acceptance rate, were on mobile devices (due to histograms and tables not displaying correctly), were not currently employed, were outside of the U.S. or using VPNs, or failed an attention check. After selection into the study, we split participants into primary and secondary samples (see below). After removing eight participants for not fully completing the exercise, the final sample sizes for the two samples were $N = 195$ (primary sample) and $N = 94$ (secondary sample).

Procedure and Conditions. After the screening and consent, participants received instructions for their task. Participants were first told that they would earn $0.05 for each correct

answer on 35 (primary sample) or 70 questions (secondary sample) for a possible bonus of $1.75 or $3.50, respectively, in order to ensure participants were motivated for each decision. Next, using instructions adapted from Reb and Cropanzano (2007) participants learned they would be acting as a Regional Supervisor to 35 sales personnel and evaluating 35 pairs of employees. Employees' performance in this organization equally relied on high mean performance and high consistency: "For this company, you care just as much about HIGHER AVERAGE performances as you do about MORE CONSISTENT performances. This is because your business model equally depends on selling many products as well as having a consistent and predictable supply chain."

Participants were randomly assigned to view employee performance data in table form or as histograms (see Supplement 8). Additionally, participants were assigned to one of two performance evaluation groups. For the primary group, participants also rated the relative overall performance of the employees and were asked to equally consider both higher mean performance and higher consistency of performance. This allowed us to test for a reduction in distribution neglect when participants judged performance via histogram compared to table data format. In the secondary group, we used the same study design but measured subjective mean and consistency of the employee pairs rather than an overall performance evaluation, allowing us to rule out the possibility that any reduction in distribution neglect is caused by more accurately identifying variance information in the histogram conditions.

Employee data consisted of weekly performance scores in dollar amounts for the previous 26 weeks. These dollar amounts represented how much more or less the employee earned for the company relative to the average employee that week. Weekly performance scores ranged from $4574 to -$4158 (M = $0; SD = 1819.24). Participants were then shown example

data (in table or histogram format, depending on their condition) and asked to answer three comprehension checks about the data. We used 35 employee profiles from Reb and Cropanzano (2007), with each profile appearing in two of the 35 pairs. Participants then rated the 35 employee pairs in counterbalanced order. Unlike Study 3's assembly-line paradigm, where rapid performance evaluations mirror real life manufacturing situations, Study 4 admittedly lacks some verisimilitude. In particular, making evaluations of 70 employees in such short order does not map on to most real-life performance evaluation settings. Rather, we use this as an internally valid paradigm to capture distribution neglect while again controlling for task difficulty, with no claim to external validity (Mook, 1983).

Measures.

Objective mean difference. We z-scored each of the 35 employee profiles and then subtracted the second employee's mean z-score from the first employee's mean z-score, creating a positive score when the first employee's mean performance was greater than the second employee's mean performance.

Objective consistency difference. Like the objective mean difference, we computed the objective consistency difference by z-scoring the standard deviations across the 35 employees and subtracting the standard deviation z-score of the second employee from the z-score of the first employee. This value was then reverse-coded, resulting in a positive difference score when the first employee was more consistent than the second employee.

Subjective performance rating difference. For each employee pair in the primary performance group, participants were asked, "Equally weighting average performance and consistency, employee X performed _____ than employee Y" with response options on a seven-point scale (1 = Much worse; 7 = Much better).

Subjective mean rating difference. For each employee pair in the secondary performance group, participants in the mean and consistency rating conditions were asked, "Employee X's AVERAGE performance is _____ than employee Y's AVERAGE performance" on a seven-point scale from 1 (much smaller) to 7 (much larger). This results in a measure that is greater when participants rate the first employee's performance as higher than the second.

Subjective consistency rating difference. For each employee pair in the secondary performance group, participants filled in the blank to this statement, "Employee X is _____ than Employee Y" using a 7-point scale from 1 (Much less consistent) to 7 (Much more consistent). This results in a measure that is greater when participants view the first employee's performance as more consistent than the second employee's performance.

##
INSERT TABLES 10 AND 11 ABOUT HERE
##

Results and Discussion

Descriptive statistics and zero-order correlations for the measures in the primary performance group for this study are shown in Table 10. To evaluate the primary prediction that displaying information in histograms reduced distribution neglect relative to displaying information in table form, we tested a set of crossed mixed-effects model using subjective performance rating difference as the dependent variable, objective mean difference and objective consistency difference as predictors, and participant and employee pair as fully-crossed grouping variables. This allowed a test of the relative relationships between the objective mean and consistency of the employee data and participants' subjective performance evaluations.

First, we evaluated these relationships in a set of two paired models, run separately in the table condition (Model 1) and then in the histogram condition (Model 2). The coefficients listed

in Table 11 show that objective consistency difference does not have a statistically significant effect on performance in the table condition (b = -0.05, SE = 0.06, p = .400), but it does in the histogram condition (b = 0.10, SE = 0.03, p = .006). We tested the statistical significance of this difference by adding a dummy variable for histogram data format along with interaction terms of that format variable with both objective mean differences and objective consistency differences (Model 3). Participants using a histogram relied significantly less on objective mean difference for performance rating (b = -0.29, SE = 0.02, $p < 0.001$) and significantly more on objective consistency difference (b = 0.15, SE = 0.02, $p < .001$; see Figure 1). In total, these results provide support for Hypothesis 4, which states that increasing completeness or salience of distributional information will increase its use in explaining performance.

Additional analysis evaluated the secondary performance group in which we measured participants' subjective evaluations of differences in employee performance averages and consistency. In line with the findings from Study 3, the correlation between objective mean difference and subjective mean difference (r = .74) is substantially larger than the correlation between objective consistency difference and subjective consistency difference (r = .30), suggesting that participants are more accurate at evaluating mean differences between the employee pairs relative to consistency differences between the same pairs. This difference is statistically significant (z = 24.912, $p < .001$), providing additional support for Hypothesis 3 which predicts that individuals will estimate distributions less accurately than averages (Lee & Preacher 2013). Further, correlations between subjective and objective ratings for both measures were higher in the table condition (mean: r = .81; consistency: r = .34) than in the histogram condition (mean: r = .65; consistency: r = .26). Thus, increased accuracy does not appear to be an alternative explanation for the reduction of distribution neglect in the histogram condition.

Study 5: Skew Neglect

In addition to variance, based on our taxonomy, skewness is another distributional characteristic that, if ignored, can lead to incomplete reasoning about the population from which it is drawn. We examine this form of distribution neglect in an experiment which also includes conditions which vary the salience of skewness as one of several plausible explanatory mechanisms for the above-average effect (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995).

Methods

Participants. We recruited 1000 participants from Prolific.co, requiring residence in the United States and fluency in English. Participants received $1.50 for the 10-minute survey. After cleaning the data for inattentive behaviors as outlined in the pre-registered analysis plan, the final sample size was N = 867 (86.7%). The median age of participants was 32 years, and 46% of the sample self-identified as female.

Procedure and conditions. Participants responded to the following prompt in an open-ended text box: “A survey of university students in the United States finds that more than half of them (about 65%) believe they are better-than-average students in terms of their grades. What might explain this? Please list any and all reasons you can think of in the order in which they come to mind. Please include all the reasons that you think are valid or relevant.” The 65% figure reflects the number of Americans who rate themselves as more intelligent than average in recent large-sample surveys (Heck, Simons, & Chabris, 2018).

Participants also viewed hypothetical histograms representing how grades might be distributed among university students in the United States (uniform, normal, skewed left, skewed right), and picked the distribution that matched how they believed grades to be distributed. The

order of these questions was determined via randomized experimental condition to vary the salience of skewed distributions. In the naïve condition participants answered the open-ended response first and then viewed the histograms on a subsequent screen. In the prompted condition, participants viewed the histograms first and then answered the open-ended response on a subsequent screen. We expected that this second condition would increase the salience of a skewed distribution, and thus reduce distribution neglect in subsequent open-ended explanations of the above average effect. The complete study materials are provided in Supplement 9.

Measures.

Free-response explanation type. Two research assistants, blind to conditions and hypotheses, coded all free responses into one of four categories: (1) egocentric bias, in which the writer assumes it is unlikely or impossible that 65% of students are actually better than average (implicitly assuming a symmetrical distribution of grades), (2) skew reasoning, in which the writer recognizes that because of skewed distributions 65% students can be better than average, (3) other, in which the writer used neither ego nor skew reasoning, and (4) multiple, in which the writer combined ego and skew reasoning in a given statement.

Statistical courses, need for cognition, mathematical and statistical proficiency. Measured as in Studies 1 and 3.

Results

Choice of histogram. To understand the baseline assumptions of participants regarding grade distributions, we first examined the descriptive statistics on histogram selection in the prompted condition (since histograms were presented first in that condition). Participants were presented with different histograms and ask to choose which one best characterized grade distributions at universities in the United States. We found that 47.1% of participants selected a

normal distribution, 2.1% selected a uniform distribution, 2.8% selected a right-skewed distribution, and 48.0% correctly selected a left-skewed distribution. There was slightly more selection of left-skewed distributions (53.9%) and less selection of normal distributions (42.5%) in the naïve condition, but the reasons for this are unclear because of the sequencing of the stimuli. Since the histograms in this condition were presented *after* the open-end response, selections could reflect a combination of baseline awareness of grade distributions or the influence of the open-end response. Overall, there was only modest evidence of skew neglect when participants were prompted with visual depictions of distributions: roughly similar percentages of participants selected a normal curve and left-skewed distribution for U.S. grades. However, the results for spontaneous free-response explanations were very different.

##

INSERT TABLES 12 AND 13 ABOUT HERE

##

Free response. Count and percentage data for the different categories of free-response reasoning are provided in Table 12. Per the preregistered analysis plan, we used an exact binomial test to evaluate the degree to which egocentric reasoning was used relative to skew reasoning. When evaluated across all conditions, participants used egocentric reasoning significantly more often than random chance (0.310, CI95%[0.289,0.332], $p < .001$) and significantly more than skew reasoning when compared pairwise (0.768, CI95%[0.735,0.798], $p < .001$). We replicate these analyses within each condition with similar results (see Table 13 for count results). These results support Hypothesis 1, which states that distribution information (in this case, distribution skew) is used less often than mean information when attempting to explain performance-related outcomes.

##

INSERT TABLE 14 ABOUT HERE

##

To evaluate the effect of salience condition on distribution neglect, we pre-registered an OLS regression model with robust standard errors clustered by participant regressing the category of the first free-response reason provided (0 = not skew reasoning, 1 = skew reasoning) on condition with the *Naïve* condition being the base case. (Note that this model, with its binary outcome, would traditionally be evaluated with a logistic regression. However, an OLS regression is suitable for these kinds of outcomes in experimental data, see Gomila, 2020). When modeled as a logistic regression, the pattern of results is almost identical, as reported in Supplement 10. This regression model evaluates whether increasing the salience of non-symmetrical distributions for participants is associated with changes in skew reasoning and therefore distribution neglect. The results of this model are shown in Table 14. Relative to the base *Naïve* condition, skew reasoning was not statistically increased by providing pictures of possible distributions (b = 0.02, SE = 0.02, p = .296). Thus, we failed to find support for Hypothesis 4, which expects that distribution neglect would be attenuated in the prompted condition.

Secondary analyses. As indicated in our pre-registration, we examined several individual difference moderators (similar to Studies 1 and 3) that may be associated with the spontaneous use of skew reasoning as well as distribution information more generally. Using OLS regression, we tested separate models to evaluate the degree to which each individual difference of interest was associated with the coded categorization of the first free-response reason provided by respondents. Need for cognition (b = 0.02, SE = 0.01, p = .168), number of

statistical courses taken (b = 0.00, SE = 0.00, p = .675), self-rated math reasoning (b = 0.01, SE = 0.00, p = .244), and self-rated statistical reasoning (b = 0.00, SE = 0.00, p = .929) were not significantly associated with the use of skew reasoning. These patterns were the same when re-analyzed with a logistic regression, as described in Supplement 10. Since these main effects were all non-significant, we did not test if they were moderated by condition.

Discussion

In this study, we began to extend our taxonomy of distribution neglect to include neglect of skew or asymmetry. In contrast to our experiments involving simulated data (see Studies 1, 3, and 4), real grade data in the United States are objectively left skewed. Participants neglected to spontaneously consider a skewed distribution for the better-than-average effect even though performance is indeed left skewed, further supporting Hypothesis 2 that distributional information (skew) is underutilized. Interestingly, many participants were able to correctly recognize that grades in the U.S. are skewed when selecting among histograms representing different distributions, but only rarely self-generated skew-based explanations. This suggests that participants are not completely unable to engage in distribution-based reasoning, but that doing so may require extra individual effort or situational encouragement, a point we return to in the General Discussion.

Unlike with variance-based reasoning (Studies 1 and 4), we found no evidence that visual presentation prompted participants to think of skew-based explanations for the above-average effect. It is possible that skew-based reasoning is even less intuitive than variance-based reasoning. Indeed, psychological scientists formally trained in statistics have routinely explained the above-average effect in terms of egocentric bias, neglecting skew-based explanations (Einhorn, 1986; Krueger & Funder, 2004). Alternatively, perhaps our manipulation of salience

via question order was insufficient to counter distribution neglect. Other interventions or manipulations could prove to be more effective.

General Discussion

The present studies provide converging evidence that although both means and distributions are important in judgment, people neglect to adequately consider variance and skew across diverse and consequential contexts. Study 1 found that individuals fail to spontaneously generate plausible variance-related causes of group differences in performance outcomes, and that this tendency was reduced but not eliminated by presenting the full distribution and summary statistics. Study 2 found that NBA managers undervalue consistent performers relative to their objective contribution to success, relying too heavily on average performance when deciding compensation without factoring in variability sufficiently, demonstrating distribution neglect in real-world performance settings. Study 3 found that people struggle to accurately assess variance differences in performance even when specifically asked to consider them, suggesting taking dispersion into account is a real problem for human decision makers. In Study 4, displaying performance scores in the form of a histogram helped reduce neglect of variance information, suggesting that presenting dispersion visually to make it simpler can help. Expanding our taxonomy of distribution neglect, participants in Study 5 failed to generate skew-related explanations for seemingly biased self-assessments of performance. At the same time, visual presentation of different distributions had no measurable effect on skew neglect, highlighting that this approach will not work in every case (Study 5).

Overall, these empirical investigations provide substantial, although not unanimous, support for our theoretical hypotheses and typology of distribution neglect. Hypothesis 1, that mean information would dominate variance and skew information, is supported by Studies 1, 2,

and 5; Hypothesis 2, underutilization of variance and skew relative to what would optimize accuracy, is supported by Studies 2 and 5; Hypothesis 3, that people are better at estimating averages than distributions, is supported by Studies 3 and 4; and Hypothesis 4, that increasing the salience of distribution information will promote its use is supported by Studies 1 and 4 but not the results of Study 5. Any of the individual studies presented here is of only limited information value in isolation, yet they collectively provide initial evidence of a multi-faceted phenomenon of distribution neglect.

Notably, the NBA study (Study 2) provides by far our strongest evidence of neglect of distributions relative to a normative benchmark, with some further evidence provided by the skew neglect study (Study 5). Further studies relying on real-world data in ecologically valid settings are needed to demonstrate that people attend to and use distribution information less that what would maximize accuracy. At the same time, further controlled experiments are needed to sample stimuli sufficiently broadly to draw general conclusions (Judd, Westfall, & Kenny, 2012; Monin & Oppenheimer, 2014; Monin, Pizarro, & Beer, 2007; Wells & Windschitl, 1999; Westfall, Judd, & Kenny, 2015). Only replicable and generalizable experimental and field evidence will allow for the strong conclusion that distribution neglect is a robust and pervasive social-cognitive phenomenon. Building on the initial studies presented here, we discuss potential boundary conditions, organizational implications, and future research directions.

Potential Interventions and Boundary Conditions

There are a number of potential boundary conditions that could moderate distribution neglect. Below we review circumstances in which we theorize distribution-based reasoning might be more (or less) likely to emerge.

Salience. Distribution neglect is neither omnipresent nor inevitable. Individuals do consider variance information in some circumstances (e.g., Kelley, 1967; Parks & Stone, 2010; Reeder & Brewer, 1979), such as when variance information is extremely salient. For example individuals consider variance more when explicitly told that an employee is consistent or not (Parks & Stone, 2010), when an entire distribution is shown at once (e.g., condition 5 of the present Study 1 and the histogram condition in Study 4), or when individuals naturalistically experience inconsistent performance in a single experimental session (Parks & Stone, 2010). Future research should further explore the role of salience in moderating distribution neglect, especially in light of the null effect of Study 5's intervention.

Computational difficulty. If distribution neglect is driven, at least in part, by the greater computational difficulty of calculating variance and skew, then this would be an important boundary condition and potential source of future interventions. Providing the full distribution of scores, summary statistics, and using histograms (Studies 1 & 4; although see Study 5) may prove effective interventions because they reduce this computational difficulty. Further consistent with this idea, the majority of the relevant experiments (Studies 1 and 3 but not Study 5) found evidence that participants high in need for cognition, who are chronically motivated to process information in depth, engage in more distribution-based reasoning. Future research should explore potential moderating factors such as cognitive load (Mitra, McNeal, & Bondell, 2017), speeded responses (Fuchs et al., 2008), rational-intuitive framing (Denes-Raj & Epstein, 1994), amount of sleep (Barnes, Jiang, & Lepak, 2016; Barnes, Lucianetti, Bhave, & Christian, 2015), and interactions between chronotype and time of day (Gunia, Barnes, & Sah, 2014).

Expertise and statistical training. Study 2 found that NBA managers underused distribution information, suggesting that domain experts can still be subject to distribution

neglect. Further, across Studies 1, 3, and 5 number of statistical courses taken and self-rated proficiency in statistics did not moderate the tendency to engage in variance-based or skew-based reasoning. That said, expertise may still moderate. Future research should track decisions makers longitudinally, as they gain more domain expertise, to see if there is any improvement over time in their use of information about distributions (Bassok, 1990; Lehman & Nisbett, 1990). Expertise can reduce the cognitive load necessary to complete computationally difficult tasks (Mitra, McNeal, & Bondell, 2017). Therefore, if experts do exhibit less distribution neglect than nonexperts, it could be because expert analyses of distributions are not as computationally challenging to them and thus expertise may interact with computational difficulty in predicting distribution neglect. Future studies exploring this idea could examine distribution neglect crossing sleep-deprivation and expertise to test for an interaction. Such manipulations may have a greater effect on how nonexperts reason about variance than on experts.

Moral vs. non-moral domain. Reeder and Brewer's (1979) schematic model of dispositional attribution distinguishes between different types of attribution structures. Hierarchically restrictive schema (e.g., performance skill) lead attributes to be interpreted as individual maxima. Performers can reach their potential but cannot perform any higher; therefore, performers with higher potential skill have a wider range of possible performance outcomes. As previously discussed with regards to the present Study 2, in a sporting context everyone can perform poorly, but only tremendous players can perform tremendously. Therefore, peak performances are attended to but weaker performances are discounted. However, in the moral domain, negative outliers are seen as highly diagnostic; therefore, people may pay particular attention to such distribution information in moral domains.

Independent vs. interdependent work. Taking performance variability into account may be more important, and occur more often, in some work contexts than in others. As discussed in Study 2, if a team's work is highly interdependent, then the output of one employee's work is the input for another employee's work. Greater variability in the quality or quantity of work in such a team would be especially problematic for the team's overall efficiency. However, in many independent work tasks, average performance may be the overwhelming consideration when it comes to assessing performance quality. Thus, managers may pay more careful attention to distributions when they supervise interdependent teams. Organizational leaders should consider and empirically examine the importance of consistency for specific employee tasks in order to ensure performance variance information is being weighted in performance evaluations appropriately given its relative importance for the organization.

Groups vs individuals. In light of the monolithic perceptions people may have of groups compared to individuals (Abelson, Dasgupta, Park, & Banaji, 1998; Dasgupta, Banaji, & Abelson, 1999), people may perceive individuals as more likely than groups to exhibit variance in their performance. Future studies could compare lay theories of dispersion to real-world performance data from both sports teams and individual players to test this idea.

Time perspective. Making predictions about longer time spans, such as years rather than weeks, may prompt individuals to anticipate variability in performance outcomes due for instance to naïve theories of trajectories of change over time (DeNisi & Stevens, 1981; Ferris et al., 2018; Reb & Cropanzano, 2007). Conversely, performances that unfold gradually, such as player statistics over the course of the year, could render variability less noticeable in situ. For example, NBA managers may have underweighted variance information in our sample because

they must observe performance and dispersion over time, rather than acquiring all the knowledge in one session.

Algorithms. Distribution neglect could potentially be reduced by relying in part on statistical algorithms, rather than solely human judgment, to make some decisions. This strategy is already being used by practitioners to overcome unrealistic optimism: Some construction firms routinely employ "optimism bias uplifts," for instance mechanically adding 30% to their planned completion times, to correct for systematic human planning biases (Flyvbjerg, 2008; Flyvbjerg, Glenting, & Rønnest, 2004). For distribution neglect, consistency of performers could be an example of a "Moneyball"-type inefficiency in the market that is corrected once it is discovered and accounted for with data analytics (Hakes & Sauer, 2006). While it can be difficult to get people to overcome the aversion to using algorithms when such algorithms make errors (Dietvorst, Simmons, & Massey, 2015), it is possible if human beings are able to make some adjustments to the algorithm (Dietvorst, Simmons, & Massey, 2016) and perhaps in other circumstances as well (Logg, Minson & Moore, 2019). Organizations of all types should consider applying data analytics to evaluating the performance of their members, and make sure to identify and include variance information in selection, promotion, and compensation decisions.

Toward a Comprehensive Model of Reasoning About Variability

We proposed and tested a taxonomy of reasoning about variability, in which some characteristics of distributions are more intuitive than others. More research is needed to further develop this hypothesized taxonomy. We propose that people typically rely on simpler approaches, with means more intuitive than variance, and normal curves more intuitive than skewed distributions. Across Studies 1-4, we find converging evidence that mean-thinking

indeed dominates variance-based reasoning. In a supplementary study reported in Online Supplement 12, we find no evidence for a hypothesized range bias such that range is considered prior to standard deviation. However, we do find evidence that people tend to neglect to consider skew in their spontaneous reasoning (Study 5). An implicit or explicit assumption of normality may be problematic for organizations that force a normal curve for performance evaluations, because real contributions follow the power law and are heavily skewed (Aguinis & O'Boyle Jr., 2014; Bersin, 2014). Artificially imposing a normal curve fails to sufficiently distinguish good performers from great performers and may lead to under-rewarding superstars for the disproportionate value they bring to the organization.

Future research should develop a more comprehensive framework for when and how people fail or succeed at factoring in distributional information. One possibility is that reasoning about means and different distributional forms is a multi-stage process. Indeed, one potential reason for the hypothesized primacy of mean-thinking is that the implications of distribution shape can be contingent on average scores. For example, whether low variance is a good thing or not (and therefore whether it should be valued and rewarded or not) is contingent on the person's average performance to some extent. If average performance is low, it might be rational to prefer more rather than less variance because it gives you a greater chance to reach a certain minimum performance threshold which many performance situations necessitate. In contrast, when average performance is quite high, variance is undesirable because there is a greater downside to that variance. Consistent with this idea, Jung and Kahn (2014) report that patients prefer hospitals with high variance in outcomes when survival rates are low rather than high. This suggests that observers may start by discerning average performance since this will help them know what to

make of variance in performance. They then may (or may not) progress to increasingly complex inferences regarding standard deviations, and normally distributed vs. skewed distributions.

A comprehensive model should also incorporate some of the contextual factors (e.g., information completeness and salience) and individual differences (e.g., need for cognition and expertise) that we have only begun to explore here. The initial evidence that need for cognition is a more robust moderator than statistical training suggests that motivation could be more important to distribution neglect than ability. Along similar lines, the results of Study 4 suggest that histograms increase the use, not comprehension, of information about variance in performance, and Study 5 finds that quite a few people accurately select skewed distributions from an array of histograms and yet fail to spontaneously generate skew-based explanations. Perhaps non-experts are capable of understanding and factoring in considerations such as variance and skew, but this requires greater cognitive effort and more encouragement than relying on averages. When individuals are not chronically or situationally driven to engage in such processing, they may tend to default towards mean thinking. Further, this psychological tendency is strong enough that even when ability and motivation are both high, as we see in the NBA study (Study 2), decision makers may still display some level of distribution neglect. Clearly people do use variance information some of the time, the question is how to encourage them to do this more often, while at the same time promoting understandings of more sophisticated forms of variance (e.g., skewed distributions).

Conclusion

The present research finds that individuals underutilize and misestimate distribution information (variance and skew) in a number of notable ways. Fair performance evaluations are important to employees being evaluated (Folger & Konovsky, 1989; Zhou & Martocchio, 2001)

and to the organization (Bettencourt & Brown, 1997; Fulford, 2005; Greenberg, 1990). Yet, our findings suggest that distribution information is underutilized when explaining patterns of performance and assessing an employee's value to the organization, unfairly undervaluing consistent performers relative to their contributions to group success. Therefore, organizations should train managers to accurately assess performance and adequately consider variability in performance. Addressing distribution neglect could also improve risk analyses processes (Beasley, Clune, & Hermanson, 2005), and help prevent harmful group stereotypes from developing (Hyde & Mertz, 2009). Acknowledging, and then proactively addressing, the subtle neglect of distributional information holds the potential to improve outcomes for both organizations and individual decision makers.

References

Abelson, R.P., Dasgupta, N., Park, J., & Banaji, M.R. (1998). Perceptions of the collective other. *Personality and Social Psychology Review, 2,* 243-250

Aguinis, H., & O'Boyle Jr, E. (2014). Star performers in twenty-first century organizations. *Personnel Psychology*, *67*(2), 313-350.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68*, 804–825. doi: 10.1037/0022-3514.68.5.804

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369-1378.

Barnes, C. M., Dang, C. T., Leavitt, K., Guarana, C. L., & Uhlmann, E. L. (2018). Archival data in micro-organizational research: A toolkit for moving to a broader set of topics. *Journal of Management*, *44*(4), 1453-1478.

Barnes, C. M., Jiang, K., & Lepak, D. P. (2016). Sabotaging the benefits of our own human capital: Work unit characteristics and sleep. *Journal of Applied Psychology*, *101*(2), 209.

Barnes, C. M., Lucianetti, L., Bhave, D. P., & Christian, M. S. (2015). "You wouldn't like me when I'm sleepy": Leaders' sleep, daily abusive supervision, and work unit engagement. *Academy of Management Journal*, *58*(5), 1419-1437.

Barnes, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance*, *20*(3), 259-274.

Barnes, C. M., Reb, J., & Ang, D. (2012). More than just the mean: Moving to a dynamic view of performance-based compensation. *Journal of Applied Psychology*, *97*(3), 711.

Bassok, M. (1990). Transfer of domain-specific problem-solving procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 522-533.

Beasley, M. S., Clune, R., & Hermanson, D. R. (2005). Enterprise risk management: An empirical analysis of factors associated with the extent of implementation. *Journal of Accounting and Public Policy*, *24*(6), 521-531.

Bersin, J. (2014, February 19). The myth of the bell curve: Look for the hyper-performers. *Forbes*. Retrieved from https://www.forbes.com/sites/joshbersin/2014/02/19/the-myth-of-the-bell-curve-look-for-the-hyper-performers/

Bettencourt, L. A., & Brown, S. W. (1997). Contact employees: Relationships among workplace fairness, job satisfaction and prosocial service behaviors. *Journal of Retailing, 73*(1), 39-61.

Cacioppo, J.T., & Petty, R.E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. doi:10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306-307.

Casella, G., & Berger, R. L. (2001). *Statistical inference*. Duxbury/Thomson Learning.

Clotfelter, C. T., & Cook, P. J. (1993). The “gambler's fallacy” in lottery play. *Management Science, 39*(12), 1521-1525.

Croson, R., & Sundali, J. (2005). The gambler’s fallacy and the hot hand: Empirical data from casinos. *Journal of Risk and Uncertainty*, *30*(3), 195-209.

Dane, E., & Pratt, M. G. (2007). Exploring intuition and its role in managerial decision making. *Academy of Management Review*, *32*(1), 33-54.

Dasgupta, N., Banaji, M.R., & Abelson, R.P. (1999). Group entitativity and group perception:

Associations between physical features and psychological judgment. *Journal of Personality and Social Psychology, 77,* 991-1003.

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, *66*(5), 819.

DeNisi, A., & Stevens, G. E. (1981). Profiles of performance, performance evaluations, and personnel decisions. *Academy of Management Journal, 24*, 592-602.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114-126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155-1170.

DuBois, C. L., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology*, *78*(2), 205.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment, 50*, 387–395.

Ferris, D. L., Reb, J., Lian, H., Sim, S., & Ang, D. (2018). What goes up must... Keep going up? Cultural differences in cognitive styles influence evaluations of dynamic performance. *Journal of Applied Psychology*, *103*(3), 347.

Fisher, M., & Keil, F. C. (2018). The binary bias: A systematic distortion in the integration of information. *Psychological Science*, *29*(11), 1846-1858.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. Mcgraw-Hill Book Company.

Flyvbjerg, B. (2008). Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, *16*(1), 3-21.

Flyvbjerg, B., Glenting, C., & Rønnest, A. (2004). Procedures for dealing with optimism bias in transport planning. *London: The British Department for Transport, Guidance Document*.

Folger, R., & Konovsky, M. A. (1989). Effects of procedural and distributive justice on reactions to pay raise decisions. *Academy of Management Journal*, *32*(1), 115-130.

Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L., & Lambert, W. (2008). Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology*, *100*(1), 30-47.

Fulford, M. D. (2005). That's not fair! The test of a model of organizational justice, job satisfaction, and organizational commitment among hotel employees. *Journal of Human Resources in Hospitality & Tourism*, *4*(1), 73-84.

Gabaix, X. (2016). Power laws in economics: An introduction. *Journal of Economic Perspectives*, *30*(1), 185-206.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20-29.

Gomila, R. (2020). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000920

Greenberg, J. (1990). Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology*, *75*(5), 561.

Greenhaus, J. H., & Parasuraman, S. (1993). Job performance attributions and career advancement prospects: An examination of gender and race effects. *Organizational Behavior and Human Decision Processes*, *55*(2), 273-297.

Gunia, B. C., Barnes, C. M., & Sah, S. (2014). The morality of larks and owls: Unethical behavior depends on chronotype as well as time of day. *Psychological Science*, *25*(12), 2272-2274.

Hakes, J. K., & Sauer, R. D. (2006). An economic evaluation of the Moneyball hypothesis. *Journal of Economic Perspectives*, *20*(3), 173-186.

Halliday, T. M., Thomas, D. M., Siu, C. O., & Allison, D. B. (2018). Failing to account for regression to the mean results in unjustified conclusions. *Journal of Women & Aging*, *30*(1), 2-4.

Heck, P.R., Simons, D.J., & Chabris, C.F. (2018) 65% of Americans believe they are above average in intelligence: Results of two nationally representative surveys. *PLoS ONE 13*(7), e0200103. https://doi.org/10.1371/journal.pone.0200103

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157-1164.

Hogarth, R. M. (2014). Deciding analytically or trusting your intuition? The advantages and disadvantages of analytic and intuitive thought. In *The Routines of Decision Making* (pp. 97-112). Psychology Press.

Hollinger, J. (2003). *Pro basketball prospectus*. Brassey's Inc.

Hollinger, J. (2005). *Pro basketball forecast*. Potomac Books, Inc.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*(5), 579.

Hyde, J., & Mertz, J. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences, 106*(22), 8801-8807.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69.

Jung, J., & Kahn, B. (2014). Perceptions of hospital safety records: Mean or variance? *ACR North American Advances.*

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292.

Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska Symposium on Motivation*. University of Nebraska Press.

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology*, *59*(4), 445.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27*(3), 313–327.

Lecerf, T., Ghisletta, P., & Jouffray, C. (2004). Intraindividual variability and level of performance in four visuo-spatial working memory tasks. *Swiss Journal of Psychology, 63*(4), 261-272.

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with no variable in common [Computer software]. Available from http://quantpsy.org.

Lehman, D. R., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology, 26*(6), 952-960.

Logg, J. M., Minson, J.A., & Moore, D.A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90-103.

Miller, S. (2019, August 16). Prediction: Redesign of performance management. Retrieved December 14, 2020, from https://www.shrm.org/resourcesandtools/hr-topics/compensation/pages/performance-management-redesigned.aspx

Mitra, R., McNeal, K. S., & Bondell, H. D. (2017). Pupillary response to complex interdependent tasks: A cognitive-load theory perspective. *Behavior Research Methods, 49*(5), 1905-1919.

Monin, B., & Oppenheimer, D.M. (2014). The limits of direct replications and the virtues of stimulus sampling [Commentary on Klein et al., 2014]. *Social Psychology, 45*, 299-300.

Monin, B., Pizarro, D., & Beer, J. (2007). Deciding vs. reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology, 11*(2), 99-111.

Mook, D. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379-387.

Moore, D. A. (2007). Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes, 102*(1), 42-58.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502-517.

Mosteller, F., & Tukey, J. W. (1977). Data analysis and regression: A second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*. Reading, MA.

Paola, M. D., & Scoppa, V. (2012). The effects of managerial turnover: Evidence from coach dismissals in Italian soccer teams. *Journal of Sports Economics, 13*(2), 152-168.

Parks, C. D., & Stone, A. B. (2010). The desire to expel unselfish members from the group. *Journal of Personality and Social Psychology, 99*(2), 303-310.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*(3), 1065-1076.

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology, 36*(4), 859-866.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29-46.

Prien, E. P., & Liske, R. E. (1962). Assessments of higher-level personnel: III. Rating criteria: A comparative analysis of supervisor ratings and incumbent self-ratings of job performance. *Personnel Psychology, 15*(2), 187–194.

Quattrone, G. A., & Jones, E. E. (1980). The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology, 38*(1), 141-152.

Reb, J., & Cropanzano, R. (2007). Evaluating dynamic performance: The influence of salient gestalt characteristics on performance ratings. *Journal of Applied Psychology, 92*(2), 490-499.

Reb, J., & Greguras, G. J. (2010). Understanding performance ratings: Dynamic performance, attributions, and rating purpose. *Journal of Applied Psychology, 95*(1), 213-220.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61-79.

Resnick, L. B. (1987). The 1987 presidential address learning in school and out. *Educational Researcher*, *16*(9), 13-54.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*(3), 482-486.

Scribner, S. (1984). Studying working intelligence. *Everyday Cognition: Its Development in Social Context*, 9-40.

SIOP. (2014). HRM Impact Award Winner. Retrieved December 14, 2020, from https://www.siop.org/Foundation/Awards/HRM-Impact-Award/Cargill

Sportrac.com. (2021). NBA Player Contracts. Retrieved February 20, 2021, from https://www.spotrac.com/nba/contracts/sort-value/all-time/limit-2000/

Sturman, M. C. (2007). The past, present, and future of dynamic performance research. In *Research in Personnel and Human Resources Management* (pp. 49-110). Emerald Group Publishing Limited.

Sunstein, C. R. (2002). Probability neglect: Emotions, worst cases, and law. *Yale Law Journal, 112*, 61.

Swaab, R.I., Schaerer, M., Anicich, E., Ronay, R., & Galinsky, A.D. (2014). The too-much-talent effect: Team interdependence determines when more talent is too much or not enough. *Psychological Science, 25*, 1581-1591.

Taylor S. E., & Brown J. D. (1988). Illusion and well-being: a social-psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling in social psychological experimentation. *Personality and Social Psychology Bulletin, 25*, 1115-1125.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*(3), 390-399.

Wheatley, G. H. (1991). Enhancing mathematics learning through imagery. *The Arithmetic Teacher*, *39*(1), 34-36.

Zhou, J., & Martocchio, J. J. (2001). Chinese and American managers' compensation award decisions: A comparative policy-capturing study. *Personnel Psychology, 54*(1), 115-145.

Tables

Table 1. Descriptive Statistics of Reasoning Response Codes

	Mean	Variance	Population	Vague	Off-Topic	Multiple
Total count	1050	63	69	5	81	363
% of total	64.4%	3.9%	4.2%	0.3%	5.0%	22.3%
First mentions	300	34	50	1	55	113
% of first mentions	54.2%	6.1%	9.0%	0.2%	9.9%	20.4%

N = 1631 responses from 553 respondents (total count); N = 553 (first mentions)

Table 2. Reasoning Response Codes by Condition

Condition	Mean	Variance	Population	Vague	Off-Topic	Multiple
Top 10 places	278	0	14	1	14	52
Top 10 & bottom 10	147	18	45	0	14	62
Full distribution	231	11	5	0	19	75
Summary stats (mean, SD)	204	21	1	1	10	108
Summary stats (SD, mean)	190	13	4	3	24	66

N = 1631 responses from 553 respondents

Table 3. Regression Results (DV = Variance Reasoning Used)

Condition	***b***	**SE**	***p***
Intercept	0.00	0.00	1.000
Top 10 & bottom 10	0.06	0.02	< .001
Full distribution	0.03	0.01	.001
Summary stats (mean, SD)	0.06	0.02	< .001
Summary stats (SD, mean)	0.04	0.01	.002

N = 1631 responses from 553 respondents
Standard errors are robust, clustered by respondent
All condition coefficients are relative to the *Top 10 places* condition

Table 4. Means, Standard Deviations, and Correlations Between Individual Level Variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Free agent	0.38	0.49									
2. Forward	0.38	0.49	-.00								
3. Center	0.33	0.47	-.02	-.55*							
4. Age	24.50	1.74	.18*	.02	-.05						
5. NBA tenure	174.59	86.22	-.21*	.03	-.07	.22*					
6. Prior salary	2099390	1889828	-.04	-.03	.02	.11*	.56*				
7. Performance trend	0.03	0.21	-.03	-.06	.02	.01	.00	.08*			
8. Performance SD	5.29	1.68	-.41*	.00	-.13*	-.06	.57*	.30*	.05		
9. Performance mean	6.98	4.48	-.46*	-.02	-.05	-.11*	.55*	.37*	.01	.85*	
10. Current salary	4167894	4212657	-.26*	.00	-.01	-.02	.48*	.37*	-.00	.48*	.53*

*$p < 0.05$; $N = 727$ players
Note. Free Agent, Forward, and Center are dummy variables.

Table 5. Means, Standard Deviations, and Correlations of Team Level Variables

Variable	*M*	*SD*	1	2
1. Players' performance SD	5.82	0.31		
2. Players' performance mean	8.95	0.97	.32*	
3. Team performance (wins)	39.70	12.93	-.11*	.62*

*$p < 0.05$; $N = 741$ teams

Table 6. Regression Results for Team Performance (Wins) Predicted by Aggregated Players' Performance

Predictor	*b*	*SE*	*t*	*b*	*SE*	*t*	*b*	*SE*	*t*
(Intercept)	41.00***	2.34	17.54	-43.46***	3.99	-10.88	30.61***	6.90	4.44
Players' performance mean				9.09***	0.39	23.55	10.13***	0.36	28.22
Players' performance SD							-14.12***	1.12	-12.60
	R^2 = .084***			R^2 = .484***			R^2 = .578***		
				ΔR^2 = .400***			ΔR^2 = .094***		

*$p < 0.05$; **$p < 0.01$, ***$p < 0.001$; $N = 741$ teams

Note. Fixed effects for season were also included in the analysis (but not in this table) to control for annual variance in wins.

Table 7. Regression Results for Player Salary Predicted by Player Performance

	Model 1			Model 2			Model 3		
Predictor	*b*	*SE*	*t*	*b*	*SE*	*t*	*b*	*SE*	*t*
(Intercept)	4341169.81*	2155486	2.01	-288645.20	2109599	-0.14	-381391.95	2148529	-0.18
Free agent	-1531088.14***	280686	-5.45	-522237.97	289417	-1.80	-520794.91	289681	-1.80
Forward	63202.13	315524	0.20	215778.34	299722	0.72	222008.39	301125	0.74
Center	199872.47	325890	0.61	409506.44	309959	1.32	423618.39	316075	1.34
Age	-164797.43*	78746	-2.09	-9257.95	76706	-0.12	-8952.49	76769	-0.12
NBA tenure	17560.34***	1953	8.99	9641.06***	2056	4.69	9499.74***	2145	4.43
Prior salary	0.21*	0.09	2.47	0.10	0.08	1.23	0.11	0.08	1.26
Performance trend	-181542.82	609935	-0.30	-22351.81	578713	-0.04	-36816.13	582452	-0.06
Performance mean				331290.01***	37304	8.88	321456.31***	56470	5.69
Performance SD							34386.73	148165	0.23
	R^2 = .377***			R^2 = .441***			R^2 = .441***		
				ΔR^2 = .063***			ΔR^2 = .000		

*p < 0.05; **p < 0.01, ***p < 0.001; N = 727 players

Note. Free Agent, Forward, and Center are dummy variables. Fixed effects for season were also included in the analysis (but not in this table) to control for annual variance in salary.

Table 8. Logistic Regression Results (DV = Decision Accuracy)

	Model 1			Model 2			Model 3		
Condition	***b***	**SE**	***p***	***b***	**SE**	***p***	***b***	**SE**	***p***
Intercept	-1.07	0.23	< .001	0.95	0.05	< .001	-1.13	0.23	< .001
Decision simplicity	2.45	0.30	< .001				2.73	0.30	< .001
Condition: Variance				-0.20	0.07	.007	-0.30	0.07	< .001

N = 5450 responses from 545 respondents

Mean is the base case for the condition variable

Robust standard errors used clustered by respondent

Table 9. Moderated Logistic Regression Results (DV = Decision Accuracy)

	Model 4			Model 5		
Condition	***b***	**SE**	***p***	***b***	**SE**	***p***
Intercept	-1.52	0.29	< .001	-1.14	0.34	< .001
Decision simplicity	2.77	0.30	< .001	2.78	0.29	< .001
Condition: Variance	-0.30	0.07	< .001	-1.02	0.36	.005
Need for cognition	0.10	0.05	.056	-0.01	0.07	.890
NFC X Condition				0.21	0.10	.042

N = 5440 responses from 544 respondents (incomplete data from one respondent)
Mean is the base case for the condition variable
Robust standard errors used clustered by respondent

Table 10. Descriptive statistics for Study 4, primary sample

Variable	*M*	*SD*	1	2	3
1. Objective mean difference	0.00	1.45			
2. Objective consistency difference	0.00	1.45	-.12*		
3. Data Format (Histogram = 2)	1.52	0.50	.00	.00	
4. Subjective performance rating difference	3.96	2.01	.78*	-.08*	.03*

*$p < 0.05$ (at the response level); N = 6825 responses from 195 respondents

Table 11. Mixed-Effects Regression Results (DV = Subjective performance rating difference)

	Model 1 (Table)			Model 2 (Histogram)			Model 3 (All)		
Condition	***b***	**SE**	***p***	***b***	**SE**	***p***	***b***	**SE**	***p***
Intercept	3.89	0.09	< .001	4.02	0.05	< .001	3.89	0.07	< .001
Objective mean difference	1.24	0.06	< .001	0.95	0.03	< .001	1.24	0.04	< .001
Objective consistency difference	-0.05	0.06	0.400	0.10	0.03	.006	-0.05	0.04	.247
Format: Histogram							0.12	0.04	.001
Obj. mean diff. X Format							-0.29	0.02	< .001
Obj. consistency diff. X Format							0.15	0.02	< .001

N = 6825 responses from 195 respondents

Table 12. Descriptive Statistics of Response Codes

	Ego	**Skew**	**Other**	**Multiple**
Total count	549	166	1050	6
% of total	31.0%	9.4 %	59.3%	0.3%
First mentions	348	89	426	4
% of first mentions	40.1%	10.3%	49.1%	0.5%

N = 1771 responses from 867 respondents (total count); 867 respondents (first mentions)

Table 13. Reasoning Response Codes by Condition

Condition	**Ego**	**Skew**	**Other**	**Multiple**
Naïve	291	80	553	3
Prompted	258	86	497	3

N = 1771 responses from 867 respondents

Table 14. Regression Results (DV = Skew Reasoning Used)

Condition	***b***	**SE**	***p***
Intercept	0.09	0.01	< .001
Prompted condition	0.02	0.02	.296

N = 1771 responses from 867 respondents
Robust standard errors used clustered by respondent
Condition coefficients are relative to the *Naive* condition

Figure 1. Interaction plots for Study 4, Model 3