Behavioral economics

Lecture 5 - **Noise**

Matej Lorko
matej.lorko@euba.sk
Student resources: www.lorko.sk

References:

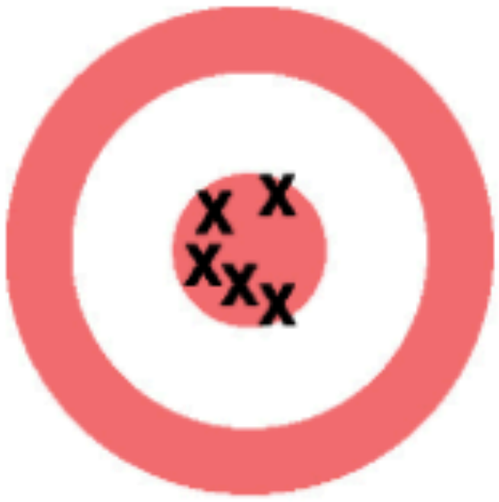
- Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: a flaw in human judgment. Hachette UK.

Judgment

- Judgment is a form of measurement in which the instrument is a human mind. Like other measurements, a judgment assigns a score to an object. The score need not be a number. “Mary Johnson’s tumor is probably benign” is a judgment, as are statements like “The national economy is very unstable,”Fred Williams would be the best person to hire as our new manager,” and “The premium to insure this risk should be $12,000.” Judgments informally integrate diverse pieces of information into an overall assessment. They are not computations, and they do not follow exact rules. A teacher uses judgment to grade an essay, but not to score a multiple-choice test.

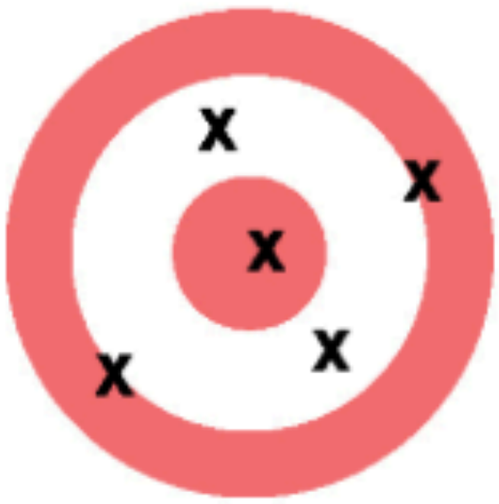
- Many people earn a living by making professional judgments, and everyone is affected by such judgments in important ways. Some judgments are predictive, and some predictive judgments are verifiable; we will eventually know whether they were accurate. This is generally the case for short-term forecasts of outcomes such as the effects of a medication, the course of a pandemic, or the results of an election. But many judgments, including long-term forecasts and answers to fictitious questions, are unverifiable. The quality of such judgments can be assessed only by the quality of the thought process that produces them. Furthermore, many judgments are not predictive but evaluative: the sentence set by a judge or the rank of a painting in a prize competition cannot easily be compared to an objective true value.

- Strikingly, however, people who make judgments behave as if a true value exists, regardless of whether it does. They think and act as if there were an invisible bull’s-eye at which to aim, one that they and others should not miss by much. The phrase judgment call implies both the possibility of disagreement and the expectation that it will be limited. Matters of judgment are characterized by an expectation of bounded disagreement. They occupy a space between matters of computation, where disagreement is not allowed, and matters of taste, where there is little expectation of agreement except in extreme cases.

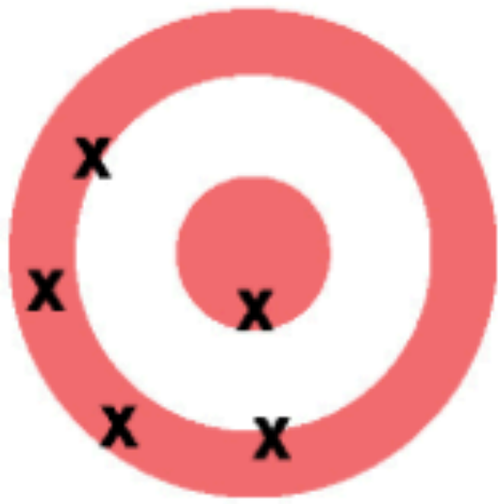


Team A

Team B



Team C



Team D

Bias and noise

- Team A shots are tightly clustered around the bull's-eye, close to a perfect pattern.

- We call Team B biased because its shots are systematically off target. As the figure illustrates, the consistency of the bias supports a prediction. If one of the team's members were to take another shot, we would bet on its landing in the same area as the first five. The consistency of the bias also invites a causal explanation: perhaps the gunsight on the team's rifle was bent.

- We call Team C noisy because its shots are widely scattered. There is no obvious bias, because the impacts are roughly centered on the bull's-eye. If one of the team's members took another shot, we would know very little about where it is likely to hit. Furthermore, no interesting hypothesis comes to mind to explain the results of Team C. We know that its members are poor shots. We do not know why they are so noisy.

- Team D is both biased and noisy. Like Team B, its shots are systematically off target; like Team C, its shots are widely scattered.

Bias and noise

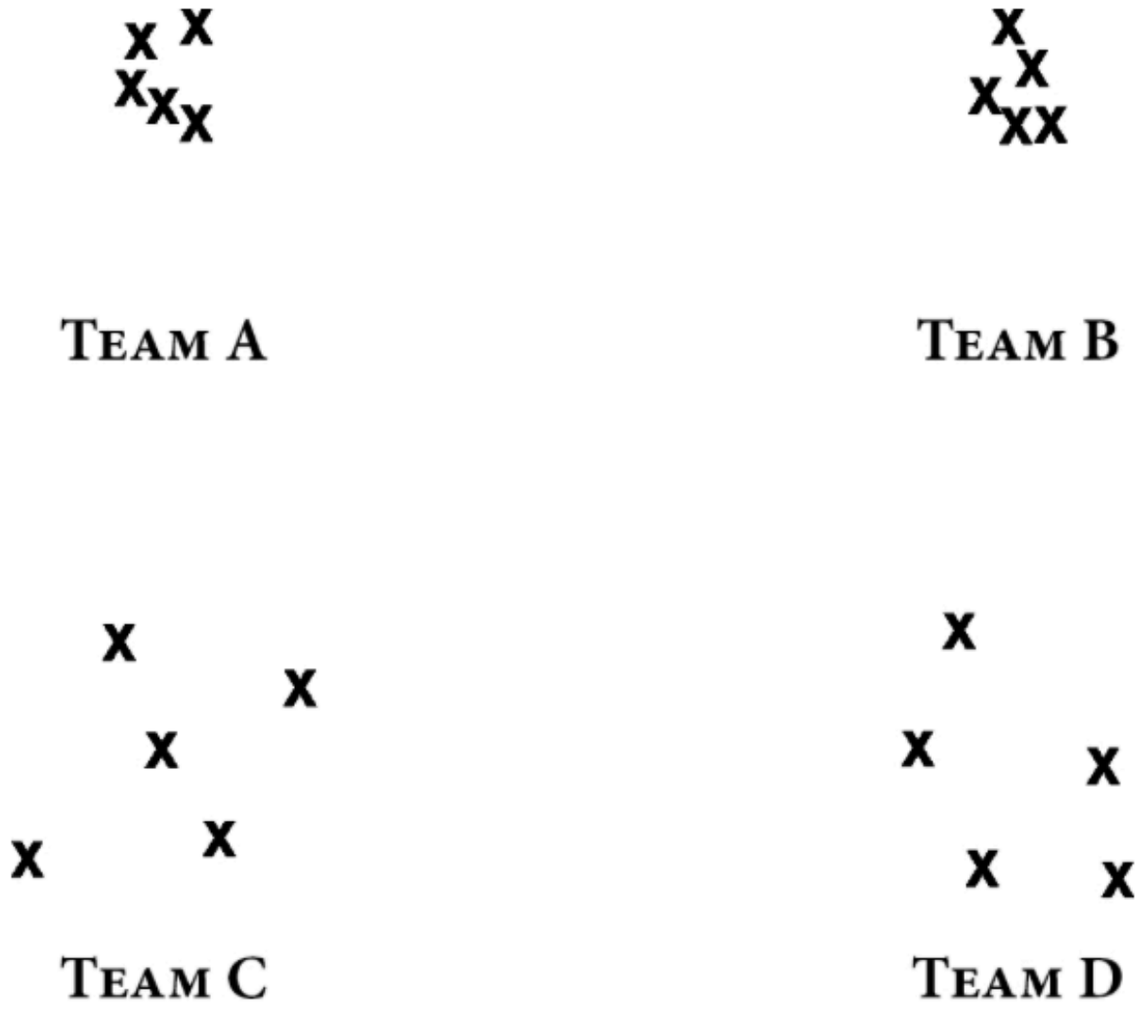


FIGURE 2: *Looking at the back of the target*

Errors: Bias and Noise

- Figure on the previous slide illustrates an important difference between bias and noise. It shows what you would see at the shooting range if you were shown only the backs of the targets at which the teams were shooting, without any indication of the bull's-eye they were aiming at.

- From the back of the target, you cannot tell whether Team A or Team B is closer to the bull's-eye. But you can tell at a glance that Teams C and D are noisy and that Teams A and B are not. A general property of noise is that you can recognize and measure it while knowing nothing about the target or bias.

- We say that bias exists when most errors in a set of judgments are in the same direction. Bias is the average error, as, for example, when a team of shooters consistently hits below and to the left of the target; when executives are too optimistic about sales, year after year; or when a company keeps reinvesting money in failing projects that it should write off.

- Eliminating bias from a set of judgments will not eliminate all error. The errors that remain when bias is removed are not shared. They are the unwanted divergence of judgments, the unreliability of the measuring instrument we apply to reality. They are noise.

- Noise is variability in judgments that should be identical. Wherever there is judgment, there is noise, and more of it than you think.

Noise

- Medicine is noisy. Faced with the same patient, different doctors make different judgments about whether patients have skin cancer, breast cancer, heart disease, tuberculosis, pneumonia, depression, and a host of other conditions. Noise is especially high in psychiatry, where subjective judgment is obviously important. However, considerable noise is also found in areas where it might not be expected, such as in the reading of X-rays.

- Child custody decisions are noisy. Case managers in child protection agencies must assess whether children are at risk of abuse and, if so, whether to place them in foster care. The system is noisy, given that some managers are much more likely than others to send a child to foster care. Years later, more of the unlucky children who have been assigned to foster care by these heavy-handed managers have poor life outcomes: higher delinquency rates, higher teen birth rates, and lower earnings.

- Forecasts are noisy. Professional forecasters offer highly variable predictions about likely sales of a new product, likely growth in the unemployment rate, the likelihood of bankruptcy for troubled companies, and just about everything else. Not only do they disagree with each other, but they also disagree with themselves. For example, when the same software developers were asked on two separate days to estimate the completion time for the same task, the hours they projected differed by 71%, on average."

- Personnel decisions are noisy. Interviewers of job candidates make widely different assessments of the same people. Performance ratings of the same employees are also highly variable and depend more on the person doing the assessment than on the performance being assessed.

- Bail decisions are noisy. Whether an accused person will be granted bail or instead sent to jail pending trial depends partly on the identity of the judge who ends up hearing the case. Some judges are far more lenient than others. Judges also differ markedly in their assessment of which defendants present the highest risk of flight or reoffending.

- Forensic science is noisy. We have been trained to think of fingerprint identification as infallible. But fingerprint examiners sometimes differ in deciding whether a print found at a crime scene matches that of a suspect. Not only do experts disagree, but the same experts sometimes make inconsistent decisions when presented with the same print on different occasions. Similar variability has been documented in other forensic science disciplines, even DNA analysis.

Noise is a problem

- Variability as such is unproblematic in some judgments, even welcome. Diversity of opinions is essential for generating ideas and options. Contrarian thinking is essential to innovation. A plurality of opinions among movie critics is a feature, not a bug. Disagreements among traders make markets. Strategy differences among competing start-ups enable markets to select the fittest. In what we call matters of judgment, however, system noise is always a problem. If two doctors give you different diagnoses, at least one of them is wrong.

- The large role of noise in error contradicts a commonly held belief that random errors do not matter, because they “cancel out.” This belief is wrong. If multiple shots are scattered around the target, it is unhelpful to say that, on average, they hit the bull’s-eye. If one candidate for a job gets a higher rating than she deserves and another gets a lower one, the wrong person may be hired. If one policy is overpriced and another is underpriced, both errors are costly to the insurance company; one makes it lose business, the other makes it lose money.

- In short, we can be sure that there is error if judgments vary for no good reason. Noise is detrimental even when judgments are not verifiable and error cannot be measured. It is unfair for similarly situated people to be treated differently, and a system in which professional judgments are seen as inconsistent loses credibility.

Measuring bias an noise

- Error in a single measurement = Bias + Noisy Error
- Overall Error (MSE) = Bias^2 + Noise^2
- The mean of squared errors (MSE) has been the standard of accuracy in scientific measurement for two hundred years. The main features of MSE are that it yields the sample mean as an unbiased estimate of the population mean, treats positive and negative errors equally, and disproportionately penalizes large errors.
- Noise in a system can be assessed by a noise audit, an experiment in which several professionals make “independent judgments of the same cases (real or fictitious). We can measure noise without knowing a true value, just as we can see, from the back of the target, the scatter of a set of shots.
- Of bias and noise, which is the larger problem? It depends on the situation. The answer might well turn out to be noise. Bias and noise make equal contributions to overall error (MSE) when the mean of errors (the bias) is equal to the standard deviations of errors (the noise).
- When the distribution of judgments is normal (the standard bell-shaped curve), the effects of bias and noise are equal when 84% of judgments are above (or below) the true value. This is a substantial bias, which will often be detectable in a professional context. When the bias is smaller than one standard deviation, noise is the bigger source of overall error.

Measuring bias an noise

	Case A	Case B	Case C		Case O	Case P	Judge Mean
Judge 1	0.5	1.5	--		13.5	12.0	**6.6**
Judge 2	2.0	--	5.5		17.5	20.0	**8.4**
Judge 3	1.5	1.8	4.0		15.0	14.0	**5.0**
Judge 207	1.0	0.5	3.0		16.0	10.0	**7.3**
Judge 208	0.5	0.3	4.0		25.5	20.0	**8.7**
Mean Sentence for Case	**1.0**	**1.1**	**3.7**		**12.2**	**15.3**	***7.0***

SD in typical case: 3.4 y.

SD of judge means: 2.4 y.

FIGURE 9: *A representation of the sentencing study*

Types of noise

- We use the term **system noise** for the noise observed in organizations that employ interchangeable professionals to make decisions, such as physicians in an emergency room, judges imposing criminal penalties, and underwriters in an insurance company.

- System noise can be broken down into **level noise** and **pattern noise.** Some judges are generally more severe than others, and others are more lenient; some forecasters are generally bullish and others bearish about market prospects; some doctors prescribe more antibiotics than others do.

- Level noise is the variability of the average judgments made by different individuals. The ambiguity of judgment scales is one of the sources of level noise. Words such as likely or numbers (e.g., “4 on a scale of 0 to 6”) mean different things to different people. Level noise is an important source of error in judgment systems and an important target for interventions aimed at noise reduction.

- System noise includes another, generally larger component. Regardless of the average level of their judgments, two judges may differ in their views of which crimes deserve the harsher sentences. Their sentencing decisions will produce a different ranking of cases. We call this variability pattern noise.

Pattern noise

- The main source of pattern noise is stable: it is the difference in the personal, idiosyncratic responses of judges to the same case. Some of these differences reflect principles or values that the individuals follow, whether consciously or not. For example, one judge might be especially severe with shoplifters and unusually lenient with traffic offenders; another might show the opposite pattern.

- Some of the underlying principles or values may be quite complex, and the judge may be unaware of them. For example, a judge could be relatively lenient toward older shoplifters without realizing it. Finally, a highly personal reaction to a particular case could also be stable. A defendant who resembles the judge's daughter might well have evoked the same feeling of sympathy, and hence leniency, on another day.

- This **stable pattern noise** reflects the uniqueness of judges: their response to cases is as individual as their personality. The subtle differences among people are often enjoyable and interesting, but the differences become problematic when professionals operate within a system that assumes consistency. Stable pattern noise that such individual differences produce is generally the single largest source of system noise.

- Still, judges' distinctive attitudes to particular cases are not perfectly stable. Pattern noise also has a transient component, called **occasion noise**. We detect this kind of noise if a radiologist assigns different diagnoses to the same image on different days or if a fingerprint examiner identifies two prints as a match on one occasion but not on another.

- As these examples illustrate, occasion noise is most easily measured when the judge does not recognize the case as one seen before. Another way to demonstrate occasion noise is to show the effect of an irrelevant feature of the context on judgments, such as when judges are more lenient after their favorite football team won, or when doctors prescribe more opioids in the afternoon.

Two candidates for an executive position

- **Monica**
- Leadership : 4
- Communication : 6
- Interpersonal skills : 4
- Technical skills : 8
- Motivation : 8
- Your prediction:

- **Nathalie**
- Leadership : 8
- Communication : 10
- Interpersonal skills : 6
- Technical skills : 7
- Motivation : 6
- Your prediction:

The Psychology of Judgment and Noise

- The judges' cognitive flaws are not the only cause of errors in predictive judgments. **Objective ignorance** often plays a larger role. Some facts are actually unknowable—how many grandchildren a baby born yesterday will have seventy years from now, or the number of a winning lottery ticket in a drawing to be held next year. Others are perhaps knowable but are not known to the judge. People's exaggerated confidence in their predictive judgment underestimates their objective ignorance as well as their biases.

- There is a limit to the accuracy of our predictions, and this limit is often quite low. Nevertheless, we are generally comfortable with our judgments. What gives us this satisfying confidence is an **internal signal**, a self-generated reward for fitting the facts and the judgment into a coherent story. Our subjective confidence in our judgments is not necessarily related to their objective accuracy.

- **Psychological biases** are, of course, a source of systematic error, or statistical bias. Less obviously, they are also a source of noise. When biases are not shared by all judges, when they are present to different degrees, and when their effects depend on extraneous circumstances, psychological biases produce noise. For instance, if half the managers who make hiring decisions are biased against women and half are biased in their favor, there will be no overall bias, but system noise will cause many hiring errors. Another example is the disproportionate effect of first impressions. This is a psychological bias, but that bias will produce occasion noise when the order in which the evidence is presented varies randomly.

- The elimination of system noise would require judges to maintain uniformity in their use of cues, in the weights they assign to cues, and in their use of the scale. Even leaving aside the random effects of occasion noise, these conditions are rarely met.

- Agreement is often fairly high in judgments on single dimensions. Different recruiters will often agree on their evaluations of which of two candidates is more charismatic or more diligent. The shared intuitive process of matching across intensity dimensionswill generally produce similar judgments. The same is true of judgments based on a small number of cues that point in the same general direction.

- Large individual differences emerge when a judgment requires the weighting of multiple, conflicting cues. Looking at the same candidate, some recruiters will give more weight to evidence of brilliance or charisma; others will be more influenced by concerns about diligence or calm under pressure. When cues are inconsistent and do not fit a coherent story, different people will inevitably give more weight to certain cues and ignore others. Pattern noise will result.

Simple models beat humans

- Most people are surprised to hear that the accuracy of their predictive judgments is not only low but also inferior to that of formulas. Even simple linear models built on limited data, or simple rules that can be sketched on the back of an envelope, consistently outperform human judges. The critical advantage of rules and models is that they are noise-free.

- As we subjectively experience it, judgment is a subtle and complex process; we have no indication that the subtlety may be mostly noise. It is difficult for us to imagine that mindless adherence to simple rules will often achieve higher accuracy than we can—but this is by now a well-established fact.

- Paul Meehl: Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence

- Lewis Goldberg: The model of you beats you

- Michael Lewis: Moneyball

- Virginia Apgar: appearance (skin color), pulse (heart rate), grimace (reflexes), activity (muscle tone), respiration (breathing rate and effort)

The Obscurity of Noise

- Cognitive biases and other emotional or motivated distortions of thinking are often used as explanations for poor judgments. Analysts invoke overconfidence, anchoring, loss aversion, availability bias, and other biases to explain decisions that turned out badly. Such bias-based explanations are satisfying, because the human mind craves causal explanations. Whenever something goes wrong, we look for a cause—and often find it. In many cases, the cause will appear to be a bias.

- Bias has a kind of explanatory charisma, which noise lacks. If we try to explain, in hindsight, why a particular decision was wrong, we will easily find bias and never find noise. Only a statistical view of the world enables us to see noise, but that view does not come naturally—we prefer causal stories. The absence of statistical thinking from our intuitions is one reason that noise receives so much less attention than bias does.

- Another reason is that professionals seldom see a need to confront noise in their own judgments and in those of their colleagues. After a period of training, professionals often make judgments on their own. Fingerprint experts, experienced underwriters, and veteran patent officers rarely take time to imagine how colleagues might disagree with them—and they spend even less time imagining how they might disagree with themselves.

- Most of the time, professionals have confidence in their own judgment. They expect that colleagues would agree with them, and they never find out whether they actually do. In most fields, a judgment may never be evaluated against a true value and will at most be subjected to vetting by another professional who is considered a respect-expert. Only occasionally will professionals be faced with a surprising disagreement, and when that happens, they will generally find reasons to view it as an isolated case. The routines of organizations also tend to ignore or suppress evidence of divergence among experts in their midst. This is understandable; from an organizational perspective, noise is an embarrassment.

Decision hygiene

- There is reason to believe that some people make better judgments than others do. Task-specific skill, intelligence, and a certain cognitive style—best described as being actively open-minded —characterize the best judges. Unsurprisingly, good judges will make few egregious mistakes. Given the multiple sources of individual differences, however, we should not expect even the best judges to be in perfect agreement on complex judgment problems. The infinite variety of backgrounds, personalities, and experiences that make each of us unique is also what makes noise inevitable.

- One strategy for reducing noise in judgment is **decision hygiene**. We chose this term because noise reduction, like health hygiene, is prevention against an unidentified enemy. Handwashing, for example, prevents unknown pathogens from entering our bodies. In the same way, decision hygiene will prevent errors without knowing what they are. Decision hygiene is as unglamorous as its name and certainly less exciting than a victorious fight against predictable biases. Noise is an invisible enemy, and preventing the assault of an invisible enemy can yield only an invisible victory. There may be no glory in preventing an unidentified harm, but it is very much worth doing.

- Six principles that define decision hygiene:

    - The goal of judgment is accuracy, not individual expression.

    - Think statistically, and take the outside view of the case.

    - Structure judgments into several independent tasks.

    - Resist premature intuitions.

    - Obtain independent judgments from multiple judges, then consider aggregating those judgments.

    - Favor relative judgments and relative scales.

The goal of judgment is accuracy, not individual expression

- Stable pattern noise is a large component of system noise and that it is a direct consequence of individual differences, of judgment personalities that lead different people to form different views of the same problem. This observation leads to a conclusion that will be as unpopular as it is inescapable: judgment is not the place to express your individuality.

- To be clear, personal values, individuality, and creativity are needed, even essential, in many phases of thinking and decision making, including the choice of goals, the formulation of novel ways to approach a problem, and the generation of options. But when it comes to making a judgment about these options, expressions of individuality are a source of noise. When the goal is accuracy and you expect others to agree with you, you should also consider what other competent judges would think if they were in your place.

- A radical application of this principle is the replacement of judgment with rules or algorithms. Algorithmic evaluation is guaranteed to eliminate noise—indeed, it is the only approach that can eliminate noise completely.

- Algorithms are already in use in many important domains, and their role is increasing. But it is unlikely that algorithms will replace human judgment in the final stage of important decisions—which is good news. However, judgment can be improved, by both the appropriate use of algorithms and the adoption of approaches that make decisions less dependent on the idiosyncrasies of one professional. Decision guidelines can help constrain the discretion of judges or promote homogeneity in the diagnoses of physicians and thus reduce noise and improve decisions.

Think statistically, and take the outside view of the case

- We say that a judge takes the outside view of a case when she considers it as a member of a reference class of similar cases rather than as a unique problem. This approach diverges from the default mode of thinking, which focuses firmly on the case at hand and embeds it in a causal story.

- When people apply their unique experiences to form a unique view of the case, the result is pattern noise. The outside view is a remedy for this problem: professionals who share the same reference class will be less noisy. In addition, the outside view often yields valuable insights.

- The outside-view principle favors the anchoring of predictions in the statistics of similar cases. It also leads to the recommendation that predictions should be moderate (regressive). Attention to the wide range of past outcomes and to their limited predictability should help decision makers calibrate their confidence in their judgments. People cannot be faulted for failing to predict the unpredictable, but they can be blamed for a lack of predictive humility.

Structure judgments into several independent tasks

- This divide-and-conquer principle is made necessary by the psychological mechanism described as excessive coherence, which causes people to distort or ignore information that does not fit a preexisting or emerging story.

- Overall accuracy suffers when impressions of distinct aspects of a case contaminate each other. For an analogy, think of what happens to the evidentiary value of a set of witnesses when they are allowed to communicate.

- People can reduce excessive coherence by breaking down the judgment problem into a series of smaller tasks. This technique is analogous to the practice of structured interviews, in which interviewers evaluate one trait at a time and score it before moving to the next one. The principle of structuring inspires diagnostic guidelines, such as the Apgar score. It is also at the heart of the approach called the mediating assessments protocol.

- This protocol “breaks down a complex judgment into multiple fact-based assessments and aims to ensure that each one is evaluated independently of the others. Whenever possible, independence is protected by assigning assessments to different teams and minimizing communication among them.

Resist premature intuitions

- The internal signal of judgment completion gives decision makers confidence in their judgment. The unwillingness of decision makers to give up this rewarding signal is a key reason for the resistance to the use of guidelines and algorithms and other rules that tie their hands.

- Decision makers clearly need to be comfortable with their eventual choice and to attain the rewarding sense of intuitive confidence. But they should not grant themselves this reward prematurely. An intuitive choice that is informed by a balanced and careful consideration of the evidence is far superior to a snap judgment. Intuition need not be banned, but it should be informed, disciplined, and delayed.

- This principle inspires our recommendation to sequence the information: professionals who make judgments should not be given information that they don't need and that could bias them, even if that information is accurate. In forensic science, for example, it is good practice to keep examiners unaware of other information about a suspect.

- Control of discussion agendas, a key element of the mediating assessments protocol, also belongs here. An efficient agenda will ensure that different aspects of the problem are considered separately and that the formation of a holistic judgment is delayed until the profile of assessments is complete.

Obtain independent judgments from multiple judges, then consider aggregating those judgments

- The requirement of independence is routinely violated in the procedures of organizations, notably in meetings in which participants' opinions are shaped by those of others. Because of cascade effects and group polarization, group discussions often increase noise.

- The simple procedure of collecting participants' judgments before the discussion both reveals the extent of noise and facilitates a constructive resolution of differences.

- Averaging independent judgments is guaranteed to reduce system noise (but not bias). A single judgment is a sample of one, drawn from the population of all possible judgments; and increasing sample size improves the precision of estimates.

- The advantage of averaging is further enhanced when judges have diverse skills and complementary judgment patterns. The average of a noisy group may end up being more accurate than a unanimous judgment.

Favor relative judgments and relative scales

- Relative judgments are less noisy than absolute ones, because our ability to categorize objects on a scale is limited, while our ability to make pairwise comparisons is much better.

- Judgment scales that call for comparisons will be less noisy than scales that require absolute judgments. For example, a case scale requires judges to locate a case on a scale that is defined by instances familiar to everyone.

Panel A

Work quality of Employee A: _____

1	2	3	4	5
Very poor	Poor	Fair	Good	Excellent

Panel B

a.

Please rate your subordinates on **safety. Safety** refers to how well the employees follow the proper rules and regulations; behave in a safe manner on the job; and demonstrate awareness and understanding of safe work practices.

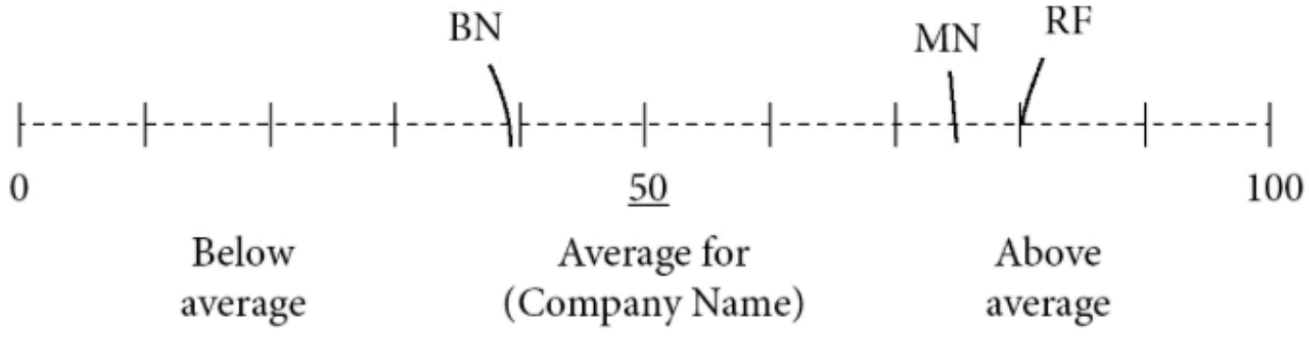


FIGURE 17: *Examples of absolute and relative rating scales*

Correcting Predictions

- 1. Make your intuitive guess.
- 2. Look for the mean.
- 3. Estimate the diagnostic value of the information you have.
- 4. Adjust from the outside view in the direction of your intuitive guess, to an extent that reflects the diagnostic value of the information you have.

- Consider, for instance, a vice president of sales who is hiring a new salesperson and has just had an interview with an absolutely outstanding candidate. Based on this strong impression, the executive estimates that the candidate should book sales of $1 million in the first year on the job—twice the mean amount achieved by new hires during their first year on the job. How could the vice president make this estimate regressive? The calculation depends on the diagnostic value of the interview. How well does a recruiting interview predict on-the-job success in this case? Based on the evidence we have reviewed, a correlation of .40 is a very generous estimate. Accordingly, a regressive estimate of the new hire's first-year sales would be, at most, $500K + ($1 million – $500K) ×.40 = $700K.

- This process is not at all intuitive. Notably, as the examples illustrate, corrected predictions will always be more conservative than intuitive ones: they will never be as extreme as intuitive predictions, but instead closer, often much closer, to the mean. If you correct your predictions, you will never bet that the tennis champion who has won ten Grand Slam titles will win another ten. Neither will you foresee that a highly successful start-up worth $1 billion will become a behemoth worth several hundred times that. Corrected predictions do not take bets on outliers.

- This means that, in hindsight, corrected predictions will inevitably result in some highly visible failures. However, prediction is not done in hindsight. You should remember that outliers are, by definition, extremely rare. The opposite error is much more frequent: when we predict that outliers will remain outliers, they generally don't, because of regression to the mean. That is why, whenever the aim is to maximize accuracy corrected predictions are superior to intuitive predictions.