ANNUAL REVIEWS

Annual Review of Organizational Psychology and Organizational Behavior

Improving Workplace Judgments by Reducing Noise: Lessons Learned from a Century of Selection Research

Scott Highhouse[1] and Margaret E. Brooks[2]

[1]Department of Psychology, Bowling Green State University, Bowling Green, Ohio, USA; email: shighho@bgsu.edu

[2]Department of Management, Bowling Green State University, Bowling Green, Ohio, USA; email: mbrooks@bgsu.edu

Annu. Rev. Organ. Psychol. Organ. Behav. 2023. 10:519–33

First published as a Review in Advance on November 28, 2022

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

https://doi.org/10.1146/annurev-orgpsych-120920-050708

Keywords

judgment and decision making, personnel selection, noise, unwanted variance, forecasting, strategic decision making

Abstract

Some assert that noise (i.e., unwanted variance) is the most neglected yet most important source of error in judgment. We suggest that this problem was discovered nearly 100 years ago in the area of personnel selection and that a century of selection research has shown that noise can be demonstrably reduced by structuring the process (i.e., decomposing the component parts, agreeing on standards, and applying those standards consistently) and by aggregating judgments independently. Algorithms can aid significantly in this process but are often confused with methods that, in their current form, can substantially increase noise in judgment (e.g., artificial intelligence and machine learning).

INTRODUCTION

Noise: unsystematic judgment deviation from a standard; the variability of error

Judge: person making a judgment

Disagreement noise: when judges disagree with one another

Occasion noise: when judges disagree with their own judgments at different time points, due to momentary state at the time of judgment

Bias: systematic judgment deviation from a standard; the mean of error

In the 1950s, Lasky et al. (1959) conducted a study examining whether experienced psychotherapists, when compared to hospital staff, made superior judgments about the future adjustment of psychiatric patients. Their main conclusion was that the ability to predict posthospital adjustment was unrelated to a staff member's professional background and training; nurses' aides and recreation therapists were just as accurate at predicting adjustment as psychologists who had regular contact with the patients. Although the failure of the psychologists to outperform the rest of the staff was the main message of the article, buried in the text was the fact that another method of judgment performed just as well—the physical weight of each patient's file as measured by a standard kitchen scale. The heaviness (i.e., thickness) of a patient's file was, when compared with the opinions of 21 different psychiatric staff members, just as good a predictor of that patient's relapse and readmission to psychiatric care.

Why did the kitchen scale perform as well as the seasoned judgments of experts? A major part of the answer is that expert judgments are noisy. In a recent book, Kahneman et al. (2021) argued that decision making researchers have ignored the unwanted variability or "noise" that necessarily decreases the accuracy of judgments.[1] In cases where consensus is desired, judges disagree with other judges (disagreement noise) and with themselves (occasion noise). Expert assessors of managerial talent, for instance, insist that they can tailor their judgments to specific circumstances, thus maximizing prediction of success within each organization (e.g., Silzer & Jeanneret 2011). Analyses of large sets of human judgments made by expert assessors, however, show that simply using a mechanical model of the assessors' policies is far more effective than a tailored assessment for forecasting managerial success (Yu & Kuncel 2022). Thus, predictions based on a model of assessor judgments across many different organizations outperform the predictions of an expert assessor making organization-specific judgments. Yet, corporations continue to pay large consulting fees for organization-specific assessments.

Purpose of the Review

In this review, we summarize how industrial-organizational (IO) psychologists have spent the past 100 years developing methods to reduce error in judgments about who will be successful. Hiring decisions affect the lives of people seeking employment, and they affect the success of organizations making these hiring decisions. Thus, judgments made in this domain can be considered high-stakes ones. Judgments made in the hiring domain are unique in that they have some standard of accuracy or success (i.e., candidate success or failure). For many workplace decisions, the decision maker may never know whether an alternative decision might have resulted in greater success. This makes employee selection a particularly useful context from which to cull strategies that might reduce noise in other domains of workplace decision making. By examining how we have developed methods for reducing noise in these kinds of judgments, we distill general principles that can generalize to less routine judgments, and those lacking a clear standard for evaluation. The review focuses on noise reduction in the selection process. We begin by reviewing literature on judgment bias, and Kahneman et al.'s (2021) recent work on noise as a judgment flaw. Next, we discuss the limits of intuition and how it adds to the problem of noise. Finally, we highlight principles that are central to noise reduction in employee selection and discuss barriers to implementing them broadly.

[1] Doherty et al. (2021) point out that unwanted variability has been a major focus of Brunswikian research and social judgment theory since the middle of the twentieth century.

JUDGMENT

System 1: intuitive, automatic system of processing that is quick and resource-conserving

System 2: logical, deliberative system of processing that is slow and resource-demanding

Judgment involves evaluating information with the goal of making inferences. It can be viewed as a psychological "weighing" of evidence (Hammond et al. 1964). In this view, judgment biases involve overweighing some pieces of evidence and underweighing, or neglecting, others (Morewedge & Kahneman 2010). Researchers have catalogued judgment biases (Kahneman et al. 1982, Nisbett & Ross 1980) estimating that there are approximately 42 (Krueger & Funder 2004), although this number may vary depending on whether a narrow or broad conceptualization of biases is adopted. In the hiring context, examples of judgment biases include a tendency to seek confirmatory rather than disconfirmatory information when interviewing (Dougherty et al. 1994), a tendency to rely too much on stereotypes in evaluating incumbents for promotion (Judge & Cable 2004, Marlowe et al. 1996), and a sensitivity to irrelevant information (Highhouse 1996).

Dual process accounts of judgment posit an intuitive system (System 1) and a deliberative one (System 2). Although there has been debate around the existence of two modes of processing in the brain (e.g., Keren 2013), the System 1 and System 2 distinction is a convenient metaphor for explaining how human judgment becomes derailed (Pennycook et al. 2018). System 1 processing is intuitive, automatic, and the default mode. System 2 processing is analytical, deliberative, and resource-demanding (Evans & Stanovich 2013). The challenge of improving judgment, therefore, is to slow down or interfere with System 1 processing. Consider this example item from the Cognitive Reflection Test (Frederick 2005):

> A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

People presented with this problem often rely on System 1 processing, giving the first response that comes to mind (i.e., 10 cents). Further reflection, or System 2 processing, reveals that a 10-cent ball would result in a total price of $1.20. The correct answer is 5 cents.

Discussions of judgmental biases in decision making often focus on the idea that people are predictably irrational—given a certain scenario, we can predict that people will behave in a specific (biased) way. It is also true, however, that individuals may incorporate different biases into their decision making, resulting in inconsistencies when multiple judges are involved. Where one judge sees applicants from inferior schools as overachievers who are likely to persevere under duress, another judge sees them as lacking the pedigree and connections to survive in a highly competitive business environment. It is therefore inconsistency in the interpretation and treatment of evidence that is pernicious for judgment.

How Is Judgment Noisy?

Kahneman et al. (2021) distinguish between biased judgments (e.g., judges are too lenient) and noisy judgments (i.e., judges are inconsistent). A mathematical way of making this distinction is to view bias as the mean of error and noise as the variability of error. Mean squared error (MSE) is composed of both bias and noise. The same MSE could be made up of high bias and low noise, or of low bias and high noise. According to Kahneman et al., the second scenario is more common in the context of human judgment. This means that judgments are often full of noise that can be reduced to enhance judgment accuracy. This is another way of saying that reliability puts a ceiling on validity, and one can improve validity (i.e., predictor—criterion linkage) merely by enhancing reliability.

Human judgment is notoriously noisy, both between and within people, such that different judges fail to agree on the relative weight to assign to different pieces of evidence, and people fail to consistently apply their own values and beliefs in practice. Kahneman et al. (2021) describe

judgment as composed of two lotteries. The first lottery concerns who is doing the judging (disagreement noise). The second lottery concerns the state of mind or mood of that judge at the time of judgment (occasion noise). Every judgment, therefore, involves unwanted variance between judges crossed with unwanted variance within judges.

Those who have studied generalizability (G) theory (Cronbach et al. 1972) should find familiar the notion that sources of noise (i.e., unwanted variance) can be isolated. G theory extends classical test theory to enable simultaneous partitioning of variance in a set of judgments into portions attributable to unique sources (e.g., across judges, within judges). It provides a kind of microscope for identifying and comparing the importance of different sources of unwanted variance in judgments. Consider the world of cat shows; owners of different cat breeds offer their pets up to be judged in the hopes that their cat will win "best of breed" or, better yet, "best of show." Moreover, serious owners travel from show to show so that they can one day boast of owning a grand champion cat. The cat show world has judges who specialize in evaluating things such as the cat's physical structure, coat, and poise. These judges travel throughout the year, often judging the same cat multiple times. This sets up an ideal situation to examine judgment variance using G theory. Variance in cat judgments could be decomposed into, for example, variance across attributes of the cats and cat breeds, variance between judges, and variance within judges across time. Moreover, one could examine the interactions of these sources of variance. Judges can (and likely do) vary wildly in their focus on a specific attribute of one specific breed—and it is also likely that they fail to agree with themselves on different days.

Although the analysis of cat show judges is hypothetical, G theory has been applied to numerous distinct kinds of workplace judgments (e.g., Greguras et al. 2003, Kraiger & Teachout 1990). These studies show that these judgments are often quite noisy and that aggregating judgments can significantly reduce that noise. Noise reduction is necessary to generalize judgments to the population of interest. One study showed, for example, that only a handful of stakeholders are needed to arrive at stable judgments of a corporation's reputation (Highhouse et al. 2009). In contrast, another study showed that an unusually large number of judges is required to establish stable population estimates of the riskiness of financial investments (Wang et al. 2022). The common theme in G theory analyses is that judges vary from one another and from themselves over time.

What Good is Intuition in Judgment?

Anecdotal examples of successful intuition are plentiful. Perhaps the most common example is that of the chess master who immediately identifies a promising move. Intuitions such as this, however, involve extensive experience and acquired skill. The conditions necessary for this kind of expertise are environments that provide high predictability and an opportunity to learn the rules (e.g., Stewart et al. 1997). Other domains have demonstrated the success of novice intuition, especially regarding aesthetic judgments (Wilson & Schooler 1991, Wilson et al. 1993). Examples of good intuition in this research stream, however, often involve anticipation of future pleasure (e.g., How much will you like this art in the future?).

Intuition is a favorite topic of researchers in the naturalistic decision making (NDM) school (Klein et al. 1993). This group typically focuses on judgments involving extreme time pressure, complexity, and crisis. Much of their evidence favoring intuition, however, involves anecdotal accounts of mission success (Lipshitz et al. 2001). According to Weiss & Shanteau (2021),

> [u]nder that [NDM] umbrella is an amalgam of methodologies and techniques, including Recognition Primed Decision Making, Situation Awareness, and Cognitive Task Analysis. The first is a finding, the second is a concept, and the third is a methodology. There is no synthesis of these ideas, no testable theory, nothing we would call scientific; there is only an opposition to the tools and ambitions of conventional JDM research. (p. 13)

The management literature on intuition similarly fails to provide testable theory and is often contaminated by hindsight bias (Dane & Pratt 2007, Salas et al. 2010).

Popular writers like Malcolm Gladwell point to psychological findings demonstrating successful intuition (e.g., Gladwell 2005). Close investigation, however, usually reveals a misinterpretation of the original research. A famous example is the perceived efficacy of "thin slicing," the notion that people can make accurate predictions based on snap judgments after limited exposure to the target. When the data are analyzed appropriately, judgments based on thin slices of interviewee behavior explain 3 to 4% of the variance in future job performance (see Eisenkraft 2013). This is similar to the most generous estimates of unstructured interviewer performance, which suggest that unstructured (i.e., traditional) employment interviews explain 4% of the variance in later job performance (Huffcutt et al. 2014).[2]

The Holistic School

The holistic school of thought, which is implicit in the use of unstructured interviews and explicit in the use of executive and special-operations assessment (Hollenbeck 2009, Prien et al. 2003, Silzer & Jeanneret 2011), is based on the idea that assessment of future success requires accounting for the "whole" person. Those who adopt the holistic viewpoint believe that expert intuition is the only way to understand how applicant attributes interact to create a complex prediction of an equally complex person. This debate predates the clinical versus actuarial debate in clinical psychology (Meehl 1954). Early IO psychologists debated the merits of generating and combining objective assessments statistically (Freyd 1925) rather than profiling and diagnosing a candidate clinically (Viteles 1925). Subsequent research has uniformly supported the statistical argument (Kuncel et al. 2013, Morris et al. 2015).

Advocates of the holistic school will often retreat to the "combined" (i.e., intuition + analytics) solution. For instance, NFL analysts condemn coaches who rely exclusively on analytics, arguing that analytics must be tempered by experience-based intuition. The addition of intuition into the hiring process, however, results in less successful predictions (Dana et al. 2013, Highhouse & Kostek 2013, Kausel et al. 2016, Kuncel et al. 2013). The earliest study to compare scores alone against scores plus assessor tinkering was conducted by T. R. Sarbin, who followed 162 incoming first-year students at the University of Minnesota in 1939. Sarbin (1943) found that admissions officers, who conducted early interviews and had access to all standardized test and high school performance data, did significantly worse in predicting first-year success than a simple (aptitude test score + high school rank) formula. Kuncel et al.'s (2013) meta-analysis showed that simple (i.e., additive) combinations of objective predictors consistently and overwhelmingly outperform predictions of assessors who have access to the same scores on the same predictors. A meta-analysis of executive assessment in the field showed that general mental ability (GMA) scores alone outpredicted holistic judgments that included the GMA scores along with other considerations (Morris et al. 2015).

The numerous problems with the assumptions behind the holistic school have been detailed elsewhere (Dawes et al. 1989, Highhouse 2008, Highhouse & Brooks 2017, Kuncel & Highhouse 2011, Ruscio 2003) and are summarized in **Table 1**. The data on this matter are clear; intuition

[2]Note, however, that criterion-related validities can be obtained only for interviews that have some sort of score associated with them. It is reasonable to assume that most unstructured interviews do not involve such a score, and those that do are likely higher in rigor. As such, the meta-analytic validity of the unstructured interview should be interpreted as a ceiling. Conway et al.'s (1995) research on the maximal reliability of job interviews supports this idea.

Table 1 Assumptions of holistic assessment and the evidence

Assumptions of holistic assessment	Evidence
Assessors can intuitively integrate substantial amounts of information in complex ways.	Simple statistical prediction rules outperform assessors in almost all cases.
Standardized tests do not predict for applicants who are educated and have reached a higher level of management.	Standardized tests predict well at the executive level and appear to be even more predictive of success in jobs of high complexity and autonomy.
Some people are better than others at making intuitive predictions of human performance.	There is no evidence of expertise in the prediction of human performance.
Holistic evaluations account for important aspects of the candidate that are ignored by decomposed methods.	Decomposed judgments outperform holistic assessments, and idiosyncratic cues are more likely to dilute than to enhance predictions.
Organizations have unique cultures, and assessments must be tailored to each job setting.	Evidence for local validities has been discredited, and the relative validity of cues remains stable across employers.

does not improve workplace judgments of talent. This does not mean, however, that there is no role for experts in the process of assessment. Experts are necessary for selecting the appropriate tools, as well as for observing behaviors and structuring the observations. It is in the combination of assessments and observations that experts run afoul.

REDUCING NOISE

Structuring Judgment

Over the past 100 years, researchers in IO psychology have found that simple things can be done to enhance the consistency, accuracy, and defensibility of judgments made in a hiring context. **Figure 1**, for instance, shows meta-analytic validities (i.e., correlations between interview scores and later on-the-job performance) for interviews at various levels of structure. Note that the predictive validity of the interview is enhanced by implementing even minimal improvements in structure. For instance, interview validity is enhanced by simply decomposing the holistic judgment and having judges make multiple evaluations along pre-established, job-related dimensions (i.e., going from the traditional unstructured interview to an interview with minimal structure).

Since the time of Meehl (1954) and Raiffa (1968), decision researchers have advocated for judges to rate targets on multiple dimensions and then amalgamate these decomposed ratings,

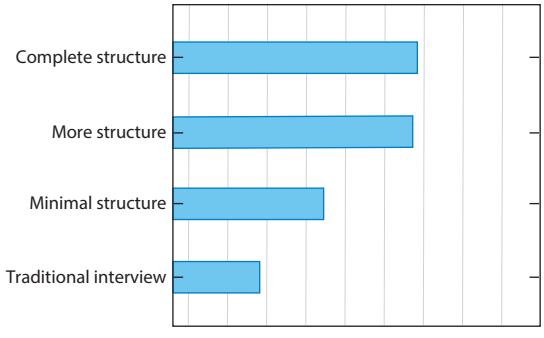

Figure 1

Validity of the employment interview at different levels of structure. Relative validities are adapted from Huffcutt & Arthur's (1994) meta-analysis of interview validities at various levels of structure for entry-level jobs.

usually by summing them (Dawes & Corrigan 1974). Arkes et al. (2006) showed that this method, when applied to ratings of presentations at a professional convention, resulted in higher inter-rater reliability (i.e., less disagreement noise) than did holistic ratings. In a follow-up study of job applicant ratings, Arkes et al. (2010) found that disaggregated ratings were more predictive of actual choices. Nevertheless, despite the higher quality of decomposed ratings, participants in Arkes et al.'s study strongly preferred the holistic approach. A general preference for inferior approaches to judgment is a theme to which we return in the section titled The Problem of User Resistance to Methods that Reduce Noise.

Figure 1 shows that a substantial further jump in interview validity occurs when the questions themselves are pre-established and scored on benchmarked answers (i.e., going from minimal structure to more structure). The highest validity can be achieved by adding the additional structure of asking applicants the exact same questions without follow-up (i.e., going to complete structure). Note that the job interview becomes increasingly predictive the more it looks like an orally administered test. This should not be surprising, as proper tests are chosen based on job-relatedness, administered in the same way to every applicant, and objectively scored. All these steps serve to reduce noise in judgments about job applicants. Tests of cognitive ability, emotional stability, and conscientiousness are predictive of success in almost any job (Barrick & Mount 2009, Sackett et al. 2022; see also the sidebar titled Getting Real About Effect Size, as well as **Figure 2**) and, contrary to popular wisdom, are useful for predicting success in extraordinarily complex jobs (Barrick & Mount 1993, Hunter 1980, Ones & Dilchert 2009).[3]

Aggregating Judgments

In a recent *Annual Review of Organizational Psychology and Organizational Behavior* article, Gigerenzer et al. (2022) claimed that aggregating judgments from multiple interviewers is counterproductive. According to this logic, if the "best" interviewer goes first, adding a second interviewer never increases accuracy. Specifically, the authors contend that "[t]he policy implication is to invest in training an excellent interviewer in each domain, increase their hit rate, and let them alone make the choice" (Gigerenzer et al. 2022, p. 185). The fundamental problem with this argument is that it assumes individual differences in interviewer accuracy, something that has yet to be demonstrated (see Pulakos et al. 1996), and seems highly unlikely given research in other domains. Moreover, research in social judgment theory (Roose & Doherty 1976) has demonstrated that composite judgments outperform even the best judge of future talent (see also Reagan-Cirincione 1994). The idea of investing in one interviewer to improve that person's hit rate, at the expense of collecting multiple judgments, is likely to create a dangerous "illusion of validity" (see Einhorn & Hogarth 1978).

In reducing system noise, Kahneman et al. (2021) emphasized that "aggregation works." Here, the authors are referring to the ability of aggregated judgments to substantially reduce noise (not bias) in the judgment process. Research on forecasting, for instance, shows that combining individual forecasts reduces errors substantially. In 30 empirical comparisons, Armstrong (2001) observed that forecasting errors were reduced by an average of 12.5% and ranged from 3 to 24% for equally weighted combined forecasts.

The most popular high-fidelity simulations for aiding hiring decisions are generally referred to as assessment centers. They include numerous individual and group managerial simulations, during which the applicants are rated on several dimensions (e.g., communication, problem solving, tolerance for stress). The combination of decomposing observer scores and observing behavior

[3] Given the lack of correlation between GMA and personality, their combined validities are often substantial (e.g., Cortina et al. 2000).

GETTING REAL ABOUT EFFECT SIZE

Bosco et al. (2015) examined nearly 8,000 effect sizes on the relation of job attitudes and behavioral outcomes. The authors found that, with rudimentary meta-analytic corrections, one can classify small (25th-percentile) effect sizes as $r = 0.07$, medium (50th-percentile) effect sizes as $r = 0.16$, and large (75th-percentile) effect sizes as $r = 0.29$. **Figure 2** applies these categories to the current state-of-the-science effect sizes for personnel selection procedures (Sackett et al. 2022). Each of the predictors shown, except for the traditional job interview, have effect sizes classified by Bosco et al. (2015, p. 443) as "especially efficacious" (i.e., predictors of behavior with effect sizes greater than $r = 0.22$, the 67th percentile). This is notable because the traditional interview is the most commonly used selection tool, and is often used to simultaneously assess candidate motivation, ability, and culture fit.

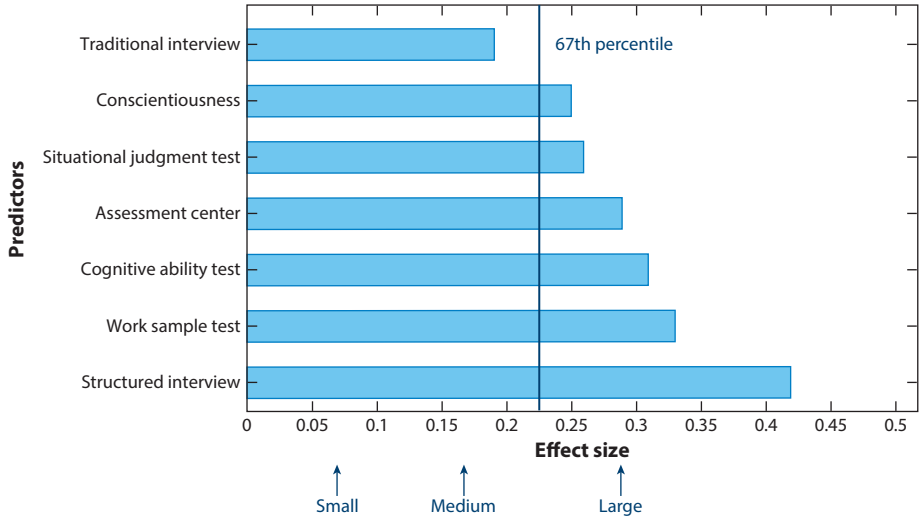

Figure 2

Effect sizes for common predictors.

across work-related simulations makes assessment centers the gold standard in managerial prediction (Sackett et al. 2017). Ironically, however, one of the assessment-center practices thought to contribute predictive validity is the assessors' consensus judgment.

A central piece of the assessment center is the overall assessment rating (OAR). The typical process involves the following: (*a*) Assessors observe and rate candidate performance in individual exercises, (*b*) they meet and report their preliminary ratings to the group for discussion, (*c*) they agree on consensus dimension ratings, and (*d*) they holistically integrate the dimension ratings to form a final OAR. Arthur et al.'s (2003) meta-analysis showed that validities of pre-consensus assessor ratings of specific dimensions by themselves (i.e., organizing and planning, problem solving, and influencing) were higher than the validity of the OAR. Dilchert & Ones (2009) conducted a study of assessment-center validity for a large sample of top-level managers, focused on the incremental validity of the OAR over and above measures of GMA and personality as measured by the Big 5. As **Table 2** shows, the OAR had zero incremental validity over tests of cognitive ability and personality. When the dimension scores were aggregated, however, the assessment center explained a substantial portion of variance not explained by GMA and personality. That is,

Table 2 Assessment center increment in validity over general mental ability (GMA) + personality for three methods of combining dimension scores[a]

Method of combination	Increment in validity over GMA and personality
Overall Assessment Rating (OAR)	0.00
Average dimension scores	0.09
Optimally weighted	0.12

[a]The data are from Dilchert & Ones (2009) two large managerial samples ($N = 4{,}985$).

the elaborate consensus rating process was detrimental to the validity of the assessment center.[4] This research is an empirical demonstration of Kahneman et al.'s (2021) dictum: "[A]ggregation works—but only if the judgments are independent" (p. 307). Assessment center practitioners have been slow to implement this important research and continue to rely heavily on suboptimal procedures (e.g., Eurich et al. 2009).

Often Ignored Problem of Applicant Noise

Although we have discussed proven processes for reducing noise in judgments about job candidates, we must acknowledge that there exists noise on the applicant side as well. Even under highly structured conditions for evaluation, the job candidates themselves add noise to the judgment process. Variance among job candidates is what makes good selection decisions a competitive advantage for organizations. Variance among job candidates that is unrelated to job success, however, constitutes additional noise in the selection system. As **Table 3** shows, candidates can have individual differences unrelated to job success that contribute to test success. These could include experience taking the kinds of assessments used by the employer, or having dispositions (e.g., surgency) that contribute to success in an interview. Other things that may cause noise on the applicant side include the applicant's momentary state (e.g., anxious, tired). The point here is that there exists a great deal of noise in the selection process, some of which seems uncontrollable. Employers must do everything possible, therefore, to reduce noise for things that are within their control. There is a critical need for research examining the effects of reducing noise on the applicant side of the job interview. We need to know whether things like providing interview questions in advance of the structured interview or allowing for repeated administration of asynchronous video interviews enhance validity by reducing interviewee noise. This idea runs counter to traditional concerns over question sharing or interview preparation as potential enemies of validity.

Table 3 Sources of job candidate noise[a]

Reasons	Example
Permanent and apply widely	A candidate is test savvy or good at outguessing items or questions.
Permanent and specific to the situation	A cheerful candidate may be more socially adept in interpersonally based assessments.
Temporary and apply widely	The candidate is tired or not feeling well.
Temporary and specific to the situation	The candidate may be slow to pick up idiosyncratic instructions.
Due to administration or scoring	There is a lack of standardization of time limit or scoring.
Chance	The candidate happens to be lucky.

[a]This table is adapted from Guion (2011).

[4]This does not imply that the consensus discussion itself is bad. Research has shown that the process of reaching consensus can result in better forecasts (Dezecache et al. 2022). The important thing is that the post-consensus ratings be made independently and combined mechanically.

The Problem of User Resistance to Methods that Reduce Noise

People who make judgments for a living want to appear competent. Methods that reduce noise in judgment undermine the appearance of competence (Arkes et al. 2007, Diab et al. 2011, Nolan et al. 2016). Arkes et al. (2007) found that physicians who used a computer-assisted diagnostic decision support system to determine the need for an ankle X-ray were consistently rated as less competent than those who did not. Similar findings occur in the employment setting, as assessors who mechanically combined scores from paper-and-pencil employment tests and a structured interview were viewed as more lazy, less personable, and less competent than assessors who holistically combined the scores (Diab et al. 2011). Research has also found that retail store supervisors assigned more relevance to candidate personality and intelligence scores when these scores were derived via a face-to-face interview than with paper-and-pencil tests (Lievens et al. 2005). It is no wonder, therefore, that professional assessors resist implementing steps to reduce noise in the system.

Structuring and decomposing judgments into dimensions (i.e., rating structured dimensions as opposed to making one holistic judgment) is also uncomfortable for judges. Employer resistance to interview structure is well-documented (Chapman & Zweig 2005, Highhouse et al. 2017, Lievens & De Paepe 2004, Nolan & Highhouse 2014, Van der Zee et al. 2002). Nolan & Highhouse (2014), for example, found that the more an interview is structured, the more an interviewer's sense of autonomy is threatened. Arkes et al. (2006, 2010) found that decomposed ratings were more reliable than holistic ratings, but that people far preferred the holistic method of judgment. The authors found some evidence to suggest that holistic ratings allowed people to factor in dimensions that are not accounted for in the decomposed approach. That is, it is easier to "game" the rating system with holistic ratings (Arkes et al. 2010). Whatever the reasons, methods that are proven to reduce noise in judgments are also methods that are uncomfortable for those making the judgments.

A quite active area of research has been focused on overcoming resistance to noise reduction strategies in employee selection (see Neumann et al. 2021 for a review). It is too early to provide recommendations for practice, but it seems clear that persuading employers to use more effective strategies is aided by providing standards against which they can compare evidence of validity (e.g., Childers et al. 2022, Highhouse et al. 2017, Zhang et al. 2018), and by enhancing user perceptions of autonomy (e.g., Dietvorst et al. 2018, Lievens & De Paepe 2004, Nolan & Highhouse 2014).

What About Algorithms?

According to Britannica (**https://www.britannica.com**), an algorithm is a "systematic procedure that produces—in a finite number of steps—the answer to a question or the solution of a problem." Under this definition, an algorithm can be an average of ratings made by one judge across a set of decomposed attributes. Alternatively, it could be an average rating of multiple assessors for an individual candidate in an assessment center. An algorithm could also be a prediction, derived from a statistical combination of predictors (e.g., ratings and tests). Algorithms are good. Algorithms reduce noise. Algorithms result in better judgments.

The problem is, however, that algorithms have come to mean many things, including artificial intelligence (AI), machine learning, and data mining (Kosinski et al. 2016). As Gigerenzer et al. (2022) noted, scraping substantial amounts of data from applicants' social media accounts for the purpose of producing machine-learning-based algorithms will result in noisy hiring judgments that are unlikely to surpass the efficacy of quite simple judgment rules (e.g., hire the candidate highest in job knowledge). Narayanan (2019), for instance, found that AI was no better than more transparent linear regression models for predicting social outcomes.

Recall that Yu & Kuncel (2022) found that a model of the typical assessor outpredicted the assessors on which the model was based. As we argued, this is because the model (i.e., an algorithm) reduced the noise that was inherent in the judgments of the executive assessors. Because algorithms are often developed to mimic humans, however, they may preserve or even amplify the systematic bias of the humans on which they were developed. Amazon famously halted development of a hiring algorithm because it penalized resumes that mentioned the word "women" (Logg 2019). Attempts to develop algorithms devoid of demographic information are extremely difficult, as proxies for this information can exist in the form of things like names and zip codes (Caliskan et al. 2017). Despite this problem, people perceive algorithms to be less biased when they anticipate being discriminated against (Jago & Laurin 2022).

Mechanical combination: combining numerical ratings of multiple dimensions or assessments based on a predetermined rule or formula

Accuracy of judgment in hiring comprises the following four steps: (1) identifying job requirements, (2) identifying the knowledge, skills, behaviors, and temperaments associated with success in completing those job requirements, (3) identifying standardized methods for assessing those things, and (4) combining those assessments in a consistently job-related way. Algorithms allow one to achieve Step 4. Unfortunately, many of the AI and machine-learning-based forms of selection too often skip Steps 1 through 3 (Tippins et al. 2021).

CONCLUSION

Good hiring is done by assessing what is knowable at the time of choice, recognizing that the outcome of the judgment can be influenced by many things outside of the employer's control. Identifying talent involves limiting the error in judgment that can be controlled. As we have argued, the largest and most controllable source of error is noise. Despite the critical role of noise in making accurate judgments, Kahneman et al. (2021) suggested that noise often goes undiscussed, often at the expense of judgmental biases. Bias, according to the authors, has a sort of "explanatory charisma" (Kahneman et al. 2021, p. 369) that is lacking for noise. It is only through a statistical view of the world, according to these authors, that we can recognize the importance of noise. This was, incidentally, the argument made in the 1925 *Journal of Applied Psychology* article, "The Statistical Viewpoint in Vocational Selection" (Freyd 1925).

As we have argued throughout this review, more than 100 years of research and practice in employee selection has demonstrated that judgment can be dramatically enhanced by structuring the process (i.e., decomposing the component parts, agreeing on standards, and applying those standards consistently) and by aggregating judgments using mechanical combination. This process can be applied to other workplace judgments that are not as frequently made and do not have some standard of accuracy or success (e.g., mergers or acquisitions). Kahneman et al. (2021), for example, use structuring a job interview as an analogy for how strategic judgments should be made: Break down the problem into multiple fact-based assessments, ensuring that each one is evaluated independently.

Eliminating noise in hiring involves a recognition that identifying the best person for the job is a probabilistic dilemma in which mistakes will be made. That professional sports teams occasionally draft duds does not invalidate the procedures used to evaluate prospects. Instead, it illustrates that even the best and most expensive methods of prediction will result in error. We are too quick to blame the limits of our technology, when uncertainty is inherent in nature (Salsburg 2001). The relevant question is how many more mistakes would be made if poorer practices were used.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Arkes HR, González-Vallejo C, Bonham AJ, Kung YH, Bailey N. 2010. Assessing the merits and faults of holistic and disaggregated judgments. *J. Behav. Decis. Mak.* 23(3):250–70

Arkes HR, Shaffer VA, Dawes RM. 2006. Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *J. Behav. Decis. Mak.* 19(5):429–39

Arkes HR, Shaffer VA, Medow MA. 2007. Patients derogate physicians who use a computer-assisted diagnostic aid. *Med. Decis. Mak.* 27(2):189–202

Armstrong JS. 2001. Combining forecasts. In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, ed. JS Armstrong, pp. 417–39. Norwell, MA: Kluwer Acad. Publ.

Arthur W Jr., Day EA, McNelly TL, Edens PS. 2003. A meta-analysis of the criterion-related validity of assessment center dimensions. *Pers. Psychol.* 56(1):125–53

Barrick MR, Mount MK. 1993. Autonomy as a moderator of the relationships between the big five personality dimensions and job performance. *J. Appl. Psychol.* 78:111–18

Barrick MR, Mount MK. 2009. Select on conscientiousness and emotional stability. In *Handbook of Principles of Organizational Behavior*, ed. EA Locke, pp. 19–40. West Sussex, UK: Wiley. 2nd ed.

Bosco FA, Aguinis H, Singh K, Field JG, Pierce CA. 2015. Correlational effect size benchmarks. *J. Appl. Psychol.* 100(2):431–49

Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain humanlike biases. *Science* 356:183–86

Chapman DS, Zweig DI. 2005. Developing a nomological network for interview structure: antecedents and consequences of the structured selection interview. *Pers. Psychol.* 58(3):673–702

Childers M, Highhouse S, Brooks ME. 2022. Apples, oranges, and ironing boards: comparative effect sizes influence lay impressions of test validity. *Int. J. Sel. Assess.* 30(2):230–35

Conway JM, Jako RA, Goodman DF. 1995. A meta-analysis of interrater and internal consistency reliability of selection interviews. *J. Appl. Psychol.* 80(5):565–79

Cortina JM, Goldstein NB, Payne SC, Davison HK, Gilliland SW. 2000. The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Pers. Psychol.* 53(2):325–51

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley

Dana J, Dawes R, Peterson N. 2013. Belief in the unstructured interview: the persistence of an illusion. *Judgm. Decis. Mak.* 8(5):512–20

Dane E, Pratt MG. 2007. Exploring intuition and its role in managerial decision making. *Acad. Manag. Rev.* 32(1):33–54

Dawes RM, Corrigan B. 1974. Linear models in decision making. *Psychol. Bull.* 81(2):95–106

Dawes RM, Faust D, Meehl PE. 1989. Clinical versus actuarial judgment. *Science* 243(4899):1668–74

Dezecache G, Dockendorff M, Ferreiro DN, Deroy O, Bahrami B. 2022. Democratic forecast: Small groups predict the future better than individuals and crowds. *J. Exp. Psychol. Appl.* 28(3):525–37

Diab DL, Pui SY, Yankelevich M, Highhouse S. 2011. Lay perceptions of selection decision aids in US and non-US samples. *Int. J. Sel. Assess.* 19(2):209–16

Dietvorst BJ, Simmons JP, Massey C. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manag. Sci.* 64(3):1155–70

Dilchert S, Ones DS. 2009. Assessment center dimensions: individual differences correlates and meta-analytic incremental validity. *Int. J. Sel. Assess.* 17(3):254–70

Doherty ME, Stewart TR, Holzworth RJ. 2021. "Noise" and social judgment theory: a commentary on Kahneman, Sibony, and Sunstein. *Brunswik Soc. Newsl.* 36:56–66

Dougherty TW, Turban DB, Callender JC. 1994. Confirming first impressions in the employment interview: a field study of interviewer behavior. *J. Appl. Psychol.* 79(5):659–65

Einhorn HJ, Hogarth RM. 1978. Confidence in judgment: persistence of the illusion of validity. *Psychol. Rev.* 85(5):395–416

Eisenkraft N. 2013. Accurate by way of aggregation: Should you trust your intuition-based first impressions? *J. Exp. Soc. Psychol.* 49(2):277–79

Eurich TL, Krause DE, Cigularov K, Thornton GC. 2009. Assessment centers: current practices in the United States. *J. Bus. Psychol.* 24(4):387

Evans JSB, Stanovich KE. 2013. Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8(3):223–41

Freyd M. 1925. The statistical viewpoint in vocational selection. *J. Appl. Psychol.* 9:349–56

Frederick S. 2005. Cognitive reflection and decision making. *J. Econ. Perspect.* 19(4):25–42

Gigerenzer G, Reb J, Luan S. 2022. Smart heuristics for individuals, teams, and organizations. *Annu. Rev. Organ. Psychol. Organ. Behav.* 9:171–98

Gladwell M. 2005. *Blink: The Power of Thinking Without Thinking*. New York: Little, Brown Co.

Greguras GJ, Robie C, Schleicher DJ, Goff M III. 2003. A field study of the effects of rating purpose on the quality of multisource ratings. *Pers. Psychol.* 56(1):1–21

Guion RM. 2011. *Assessment, Measurement, and Prediction for Personnel Decisions*. Abington-on-Thames: Routledge

Hammond KR, Hursch CJ, Todd FJ. 1964. Analyzing the components of clinical inference. *Psychol. Rev.* 71(6):438–56

Highhouse S. 1996. Context-dependent selection: the effects of decoy and phantom job candidates. *Organ. Behav. Hum. Decis. Process.* 65(1):68–76

Highhouse S. 2008. Stubborn reliance on intuition and subjectivity in employee selection. *Ind. Organ. Psychol.* 1:333–42

Highhouse S, Broadfoot A, Yugo JE, Devendorf SA. 2009. Examining corporate reputation judgments with generalizability theory. *J. Appl. Psychol.* 94(3):782–89

Highhouse S, Brooks ME. 2017. Straight talk about selecting for upper management. In *The Oxford Handbook of Talent Management*, ed. DG Collings, K Mellahi, WF Cascio, pp. 268–80. Oxford, UK: Oxford Univ. Press

Highhouse S, Brooks ME, Nesnidol S, Sim S. 2017. Is a .51 validity coefficient good? Value sensitivity for interview validity. *Int. J. Sel. Assess.* 25(4):383–89

Highhouse S, Kostek JA. 2013. Holistic assessment for selection and placement. *APA Handbook of Testing and Assessment in Psychology*. Washington, DC: Am. Psychol. Assoc.

Hollenbeck GP. 2009. Executive selection—What's right. . .and what's wrong. *Ind. Organ. Psychol.* 2:130–43

Huffcutt AI, Arthur W. 1994. Hunter and Hunter 1984 revisited: interview validity for entry-level jobs. *J. Appl. Psychol.* 79(2):184–90

Huffcutt AI, Culbertson SS, Weyhrauch WS. 2014. Moving forward indirectly: reanalyzing the validity of employment interviews with indirect range restriction methodology. *Int. J. Sel. Assess.* 22(3):297–309

Hunter JE. 1980. *Validity Generalization for 12,000 Jobs: An Application of Synthetic Validity and Validity Generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: U.S. Dep. Labor, Empl. Serv.

Jago AS, Laurin K. 2022. Assumptions about algorithms' capacity for discrimination. *Personal. Soc. Psychol. Bull.* 48(4):582–95

Judge TA, Cable DM. 2004. The effect of physical height on workplace success and income: preliminary test of a theoretical model. *J. Appl. Psychol.* 89(3):428–41

Kahneman D, Sibony O, Sunstein CR. 2021. *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark

Kahneman D, Slovic SP, Slovic P, Tversky A, eds. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge Univ. Press

Kausel EE, Culbertson SS, Madrid HP. 2016. Overconfidence in personnel selection: when and why unstructured interview information can hurt hiring decisions. *Organ. Behav. Hum. Decis. Process.* 137:27–44

Keren G. 2013. A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich 2013. *Perspect. Psychol. Sci.* 8(3):257–62

Klein GA, Orasanu J, Calderwood R, Zsambok CE, eds. 1993. *Decision Making in Action: Models and Methods*. Norwood, NJ: Ablex Publ. Corp.

Kosinski M, Wang Y, Lakkaraju H, Leskovec J. 2016. Mining big data to extract patterns and predict real-life outcomes. *Psychol. Methods* 21(4):493–506

Kraiger K, Teachout MS. 1990. Generalizability theory as construct-related evidence of the validity of job performance ratings. *Hum. Perform.* 3(1):19–35

Krueger JI, Funder DC. 2004. Towards a balanced social psychology: causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behav. Brain Sci.* 27(3):313–27

Kuncel NR, Highhouse S. 2011. Complex predictions and assessor mystique. *Ind. Organ. Psychol.* 4:302–306

Kuncel NR, Klieger DM, Connelly BS, Ones DS. 2013. Mechanical versus clinical data combination in selection and admissions decisions: a meta-analysis. *J. Appl. Psychol.* 98:1060–72

Lasky JJ, Hover GL, Smith PA, Bostian DW, Duffendack SC, Nord CL. 1959. Post-hospital adjustment as predicted by psychiatric patients and by their staff. *J. Consult. Psychol.* 23(3):213–18

Lievens F, De Paepe A. 2004. An empirical investigation of interviewer-related factors that discourage the use of high structure interviews. *J. Organ. Behav.* 25:29–46

Lievens F, Highhouse S, De Corte W. 2005. The importance of traits and abilities in supervisors' hirability decisions as a function of method of assessment. *J. Occup. Organ. Psychol.* 78(3):453–70

Lipshitz R, Klein G, Orasanu J, Salas E. 2001. Taking stock of naturalistic decision making. *J. Behav. Decis. Mak.* 14(5):331–52

Logg JM. 2019. Using algorithms to understand the biases in your organization. *Harvard Business Review*, Aug. 9. **https://hbr.org/2019/08/using-algorithms-to-understand-the-biases-in-your-organization**

Marlowe CM, Schneider SL, Nelson CE. 1996. Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *J. Appl. Psychol.* 81(1):11–21

Meehl PE. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: Univ. Minn.

Morewedge CK, Kahneman D. 2010. Associative processes in intuitive judgment. *Trends Cogn. Sci.* 14(10):435–40

Morris SB, Daisley RL, Wheeler M, Boyer P. 2015. A meta-analysis of the relationship between individual assessments and job performance. *J. Appl. Psychol.* 100(1):5–20

Narayanan A. 2019. *How to recognize AI snake oil*. Presented at the Arthur Miller Lecture on Science and Ethics, Cambridge, MA, April 22. **https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf**

Neumann M, Niessen ASM, Meijer RR. 2021. Implementing evidence-based assessment and selection in organizations: a review and an agenda for future research. *Organ. Psychol. Rev.* 11(3):205–39

Nisbett RE, Ross L. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Hoboken, NJ: Prentice Hall

Nolan KP, Carter NT, Dalal DK. 2016. Threat of technological unemployment: Are hiring managers discounted for using standardized employee selection practices? *Pers. Assess. Decis.* 2(1):30–47

Nolan KP, Highhouse S. 2014. Need for autonomy and resistance to standardized employee selection practices. *Hum. Perform.* 27(4):328–46

Ones DS, Dilchert S. 2009. How special are executives? How special should executive selection be? Observations and recommendations. *Ind. Organ. Psychol.* 2:163–70

Pennycook G, De Neys W, Evans JSB, Stanovich KE, Thompson VA. 2018. The mythical dual-process typology. *Trends Cogn. Sci.* 22(8):667–68

Prien EP, Schippmann JS, Prien KO. 2003. *Individual Assessment: As Practiced in Industry and Consulting*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Pulakos ED, Schmitt N, Whitney D, Smith M. 1996. Individual differences in interviewer ratings: the impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Pers. Psychol.* 49(1):85–102

Raiffa H. 1968. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Boston: Addison-Wesley

Reagan-Cirincione PA. 1994. Improving the accuracy of group judgment: a process intervention combining group facilitation, social judgment analysis, and information technology. *Organ. Behav. Hum. Decis. Process.* 58:246–70

Roose JE, Doherty M. 1976. Judgment theory applied to the selection of life insurance salesmen. *Organ. Behav. Hum. Decis. Process.* 22:193–15

Ruscio J. 2003. Holistic judgment in clinical practice. *Sci. Rev. Mental Health Pract.* 2:38–48

Sackett PR, Shewach OR, Keiser HN. 2017. Assessment centers versus cognitive ability tests: challenging the conventional wisdom on criterion-related validity. *J. Appl. Psychol.* 102(10):1435–47
Sackett PR, Zhang C, Berry CM, Lievens F. 2022. Revisiting meta-analytic estimates of validity in personnel selection: addressing systematic overcorrection for restriction of range. *J. Appl. Psychol.* 107(11):2040–68
Salas E, Rosen MA, DiazGranados D. 2010. Expertise-based intuition and decision making in organizations. *J. Manag.* 36(4):941–73
Salsburg D. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: Macmillan
Sarbin TL. 1943. A contribution to the study of actuarial and individual methods of prediction. *Am. J. Sociol.* 48:598–602
Silzer R, Jeanneret R. 2011. Individual psychological assessment: a practice and science in search of common ground. *Ind. Organ. Psychol.* 4(3):270–96
Stewart TR, Roebber PJ, Bosart LF. 1997. The importance of the task in analyzing expert judgment. *Organ. Behav. Hum. Decis. Process.* 69(3):205–19
Tippins NT, Oswald FL, McPhail SM. 2021. Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action. *Pers. Assess. Decis.* 7(2):1
Van der Zee KI, Bakker AB, Bakker P. 2002. Why are structured interviews so rarely used in personnel selection? *J. Appl. Psychol.* 87(1):176–84
Viteles MS. 1925. The clinical viewpoint in vocational selection. *J. Appl. Psychol.* 9:131–38
Wang LY, Highhouse S, Brooks ME. 2022. Culture versus other sources of variance in risk and benefit perceptions: a comparison of Japan and the United States. *J. Behav. Decis. Mak.* 35(5):e2277
Weiss DJ, Shanteau J. 2021. The futility of decision making research. *Stud. Hist. Philos. Sci. A.* 90:10–14
Wilson TD, Lisle DJ, Schooler JW, Hodges SD, Klaaren KJ, LaFleur SJ. 1993. Introspecting about reasons can reduce post-choice satisfaction. *Personal. Soc. Psychol. Bull.* 19(3):331–39
Wilson TD, Schooler JW. 1991. Thinking too much: introspection can reduce the quality of preferences and decisions. *J. Pers. Soc. Psychol.* 60(2):181–92
Yu MC, Kuncel NR. 2022. Testing the value of expert insight: comparing local versus general expert judgment models. *Int. J. Sel. Assess.* 30(2):202–15
Zhang DC, Highhouse S, Brooks ME, Zhang Y. 2018. Communicating the validity of structured job interviews with graphical visual aids. *Int. J. Sel. Assess.* 26(2–4):93–108