LUND
UNIVERSITY

Department of Psychology

When judging candidate's, less is more: How nondiagnostic information affects judgment in the process of personnel selection

Vilma Seth

Master's thesis (15 hp)
Spring 2022

Supervisor: Martin Bäckström

Abstract

Previous studies have found what is referred to as a *dilution effect* of nondiagnostic information when subjects make evaluative judgment tasks of future performance, meaning that the presence of nondiagnostic information makes predictions less extreme. However, a large study recently failed to replicate this effect. The hypothesis of this study is that judges can discard from obviously irrelevant information but fail to do so when the information is *seemingly* relevant (meaning apparently useful, but without an effect on performance), and that this will influence both the ranking of the candidates and the amount of variance, or noise, in the evaluations. To measure this the participants (N = 145*)* were randomly assigned to three different conditions where they judged the same 15 job candidates. In one condition only diagnostic information about the candidates was shared, in the second the participants also received obviously nondiagnostic information (about hobbies, siblings, favourite food and similar) and in the third condition the participants received diagnostic information and seemingly diagnostic information, in the form of personality traits without relation to performance and the candidate's self-description. The study found that while there was no effect of the nondiagnostic information on judgment, there was a significant effect of the seemingly diagnostic information on the ranking of candidates, as well as on the amount of variance, in judgments of individual candidates. The conclusion is that judges seem able to disregard obviously nondiagnostic information, but are still affected by seemingly diagnostic information, which affects ranking and adds variability.

Keywords: *dilution effect, personnel selection, recruiting, nondiagnostic information, noise*

Acknowledgements

I want to direct my biggest and warmest thank to my supervisor, Martin Bäckström, who tirelessly have been discussing my various research topics with me, then the experiment design, then the results and the confusing and marvelous laws of statistics. Thank you for your brilliance and your patience. I also want to thank Jenny Heffler and Viktoria Asp at Meritmind for giving me valuable insights into the world of recruiting.

When judging candidate's, less is more:

How nondiagnostic information affects judgment in the process of personnel selection

When people make evaluative judgments, they tend to combine available information and weigh them against each other to form a judgment or prediction (Brunswik, 1952). However, an impressive body of research, perhaps most notably the work of Tversky and Kahneman, tells us that humans unfortunately are quite unreliable decision makers (Tversky & Kahneman, 1973) and that we are painfully ignorant of our own shortcomings (Kahneman, 2011). This deficit of the human mind has many implications for us as a species, but for the purpose of this study we will focus on the effects it has on personnel selection.

Finding and attracting talent is considered one of the most important success factor for organizations today (Boston Consulting Group, 2007). And attaining talent is expensive, a 2016 report from SHRM (the Society for Human Resource Management) states that median cost per non-executive hire is $1633, while the average cost is $4425 (N = 488) (SHRM, 2016). A German study calculated the cost to an average of 8 weeks wages, roughly € 4700 (Muehlemann & Pfeifer, 2016). But part from the cost of hire, there is additional costs to making the wrong hiring decision, that can widely overshadow the direct costs: costs of turnover, low efficiency cost, and lost revenue from the not-recruited better fit employee (O'Connell & Kung, 2007). Conclusively there is a clear incentive for organizations to use a process for personnel selection that is both cost effective and that, at the end of the process, generates the very best candidate for the job. Today the most widely used method for personnel selection is the interview (SHRM, 2016), but this method has proven to be a very poor predictor of job performance (Kausel et al., 2016; Kuncel et al., 2014; McDaniel et al., 1994; Schmidt & Hunter, 1998). But even when informed that interviews have very low predictive value, human resource executives prefer them to a more structured approach (Highhouse, 2008; Kuncel, 2014).

A large metastudy on hiring and admission found that using mechanical (e.g., algorithms) instead of clinical (holistic human judgment) data combination improved the predictive value of decisions with roughly 50% (Kuncel, 2014). Not only is using a structured mechanical approach to personnel selection simply better at picking candidates, it is also a lot less expensive (Momin & Mishra, 2015).

Part from the perspective of the employer to get the best candidate at a reasonable price there is also an obvious need from the candidate's point of view to be judged fairly in a recruitment

process. As a candidate you should be able to have faith that you weren't offered the position you applied for due to another candidate being better, and not due to personal taste or current mood of the recruiter. This is however not necessarily the case when you are evaluated by a human being, and especially not if the evaluation is based on unstructured data (such as an interview or a cover letter) (Highhouse, 2008; Kahneman, 2021; Kuncel, 2013).

On the subject of fairness it is important to make a distinction between candidates being systematically mistreated in hiring situations due to stereotype bias, and the irregular and non-systematic effect of individual human judgment. This study aims to investigate unwanted (and unsystematic) variability in judgment, and not the systematic error in the form of bias.

Personnel selection

The field of personnel selection is one of the most researched in all of psychology (Kahneman et al., 2021). As specialisation and complexity in the world of work has increased, the need to develop instruments for finding, sorting and assessing talent in job candidates has increased (Chambers et al., 1998). We no longer need just healthy strong workers, they also need to be a good fit for both the hiring organization and the specific position to be filled (Breaugh & Starke 2000).

In order to easier assess candidates, recruiters and hiring managers use an array of different instruments for candidate evaluation. Below we will account for some of them later referred to in this study.

Instruments for personnel selection

General mental ability

General mental ability (GMA, also known as IQ) has time and again proved to be the best overall predictor for job performance (Kahneman et al., 2021; Ree et al., 1994, Sternberg et al., 2001). As formulated in one review "GMA predicts both occupational level attained and performance within one's chosen occupation and does so better than any other ability, trait, or disposition and better than job experience" (Schmidt & Hunter, 2004). IQ tests have been rightfully criticised for disadvantaging marginalised groups, especially African Americans (Onwuegbuzie & Daley, 2001). But that merely means that there could (and should) be other ways to measure IQ that better captures the intelligence of other ethnic and social groups, it doesn't take away the fact that those who do score high on traditional tests also tend to fare better in their professional life.

The fact that so many hiring managers fail to use intelligence testing as their main research tool is a sign of a big gap between research and practice (Schmidt & Oh, 2021).

Big 5 Personality Inventories

Today there is an overall agreement in the field of personality psychology around the five factor model, also known as the Big 5, which measures five distinguishable and reliable factors that constitutes personality: Openness to experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism (Barrick & Mount, 1991). But the use of personality testing in the field of personnel selection has been widely disputed, and is still so today (He et al., 2019; Morgeson et al., 2007; Rothstein & Goffin, 2006; Tett & Christiansen, 2007). Over 50 years ago Guion and Gottier (1965) concluded that there was no support for the use of personality testing in employment decision processes. This led to a mellow interest in the field for another 20 years, but in 1991 two large meta-studies sparked another wave of interest (Morgeson et al., 2007). One of them merely stated that there was "some grounds for optimism concerning the use of personality measures in employee selection" (Tett & Rothstein, 1991, p. 703) while the other concluded that out of the five personality traits conscientiousness is a near-universal predictor of job performance, regardless of type of job or industry, while extraversion and openness to experience could matter depending on the measured ability (job proficiency, training proficiency and personnel data) or vocation/position in question (Barrick & Mount, 1991). Employees with higher levels of conscientiousness are more likely to remain better performers because of their dependability, reliability, trustworthiness, and inclination to adhere to company norms, rules, and values (Levy et al., 2011). This seems to be especially true for accountants. Among students in accounting a higher level of conscientiousness predicted better performance than students with lower levels of conscientiousness (Perlow & Kopp, 2004). Another study showed that the level of conscientiousness among accountants are significantly higher than in the general population and that level of conscientiousness also predicted job satisfaction (Levy et al., 2011).

Situational judgment tests

Situational judgment tests (SJTs) are assessment methods that first present respondents with realistic work situations and then require them to identify how they would respond to each situation (Campion et al., 2014). SJTs are intended to evaluate constructs (knowledge, skills,

abilities, and other characteristics) related to job performance but that differ from those measured through cognitive ability tests or personality inventories (Sorrel et al., 2016). SJTs are scored by comparing the option selected by the respondent to a scoring key, usually developed by subject matter experts (SMEs) (MacKenzie et al., 2009). SJTs have been proven to be a valid predictor of overall job performance (Chan & Schmitt, 2002; McDaniel et al., 2007).

Self-descriptions

Self-description is better regarded as a self-presentation (meaning an instruction on how you want to be perceived) than self-disclosure (meaning revealing factual information about oneself) (Johnson, 1981). Not very surprisingly, in a job interview, where there is an apparent benefit of describing oneself in a desirable manner, self-description is poorly correlated to later job performance (Barrick et al., 2009).

Noise – the unwanted variability in judgment

Noise refers to the unwanted variability in judgment, especially focused on professional decision making (Kahneman et al., 2021). When humans make judgements, these judgements are affected by several things, like our beliefs and history, current mood and tastes (Kahneman et al., 2021). As explained in the introduction judgments that error in a shared pattern is called bias. In a recruiting situation an example of this is that we can predict that a certain group of candidates will be judged more favourably or more harshly than another. Bias in the form of discrimination can be based on stereotypes about gender (Gorman, 2005), weight (Agerström & Rooth, 2011) and race (Neckerman & Kirschenman (1991). However, the judgement of candidates also varies for other reasons, not related to bias but to the individual judge, also known as inter-rater effect or inter-rater reliability. If the hiring part sees themselves in the candidate, by sharing hobbies or leisure times pursuit, they are more likely to perceive them as a good fit for the organization (Rivera, 2012). This is labelled by Kahneman et al. (2021) as pattern noise. Other influencers can be current mood or order effect (e.g. the candidate is the last interview after a long and tiresome day for the recruiter). In their book Kahneman et al. (2021) refers to this as the second lottery, the first being who you get to be judged by. This second lottery is labelled occasion noise, which can be compared to lack of test-retest reliability.

There are many sources of noise, in this study the focus will be on how irrelevant (referred to as nondiagnostic) information creates noise in the process of personnel assessment and selection. As noise equals variability in individual judgments it is measured by comparing standard deviations on single cases (how much disagreement/variation there is in the judgment of case/candidate X). The larger the standard deviation the more noise there is.

A comment on noise theory

It is important to note that the work of Kahneman et al (2021) is not primarily based on new empirical research but is perhaps best understood as a pedagogical re-packaging of an already well-studied phenomenon. The book is also referring to a great deal of historic studies carried out by other researchers, who wouldn't say that they researched "noise". Noise equals nothing other than variability in human judgment, and the book is not to be understood as a new theory. Rather it is an umbrella term for a diverse and well-researched phenomenon, namely (unwanted or unexplained) variability in human judgment.

Dilution theory

In its essence dilution theory claims that nondiagnostic information weakens the implications of diagnostic information. Information is diagnostic if by using it the judge reach a more accurate or better decision (Dalal et al., 2020). The theory is attributed to Nisbett and Zukier who, in a series of famous studies (1981), tried to evaluate how people combine information they believe to be diagnostic (meaning relevant and helpful to predict an outcome) with information they believe to be nondiagnostic (to have no or little predictive value). It is based on research by Tversky (1977) and his features of similarity model. The model theorizes that added nondiagnostic information reduces the similarity between the stimulus and the predicted outcome. The theory of Nisbett and Zukier was that nondiagnostic information combined with diagnostic information would make predictions less extreme, would in other words *dilute* (≈ thin out, water down) judgment. In the first study they compared results between groups receiving only diagnostic information or added nondiagnostic information. The diagnostic-only groups received diagnostic info (in this case a stereotyped category label, namely what subject they majored in) either about a group of students or about one individual student, studying either music or engineering. In the third condition the participants got to see a taped interview with four students also containing "background information" judged by pre-test subjects to be nondiagnostic. The subjects were

students enrolled in an introductory psychology class (Nisbett et al., 1981). The idea behind this layout of the experiment, later replicated by several scientists, is that subjects get some information that is strongly associated to an outcome (e.g., someone is described as strong, and then you predict that they go to the gym a lot). Then you provide additional information that is not necessarily the stereotype of an avid gymgoer, maybe that the person is a catholic, or studies law. With more information added the subject is expected to perceive a loss in similarity between the subject described and the strong gymgoer, and hence they get more uncertain and less extreme in their prediction (what do I really know about this person? They might think). Although, in fact, being strong is correlated to your gym-habits while religion or studies aren't, the additional information reduces the similarity between the person described and the idea of the outcome (the heuristic example of a gym-goer, see Tversky, 1977).

One thing worth noting on the first study mentioned here by Nisbett et al. (1981) is that both diagnostic and nondiagnostic information was valued as such based on nothing else than stereotypes (diagnostic information) and pre-test subject's ratings of what type of information are diagnostic or not (nondiagnostic information was picked as such based on these ratings). So that all information rated as nondiagnostic by pre-test subjects was then also categorized as such, but they were students and not experts in the field. In one condition the subjects were asked to evaluate if music majors or engineering majors would be better at enduring pain in a shock experiment, in the other the subjects predicted movies seen in a year by English majors versus premedical majors. It might be that English majors have a larger average movie consumption than premedical majors, or that engineering majors are in fact better at taking pain than music majors, however none of these two statements were actually known, but was nevertheless rated as diagnostic information, based simply on stereotypes. The assumption that music majors would be worse at enduring pain than engineering majors is problematic, as it is based on false premises about masculinity (being supposedly higher in the group of engineers, thus predicting higher pain endurance than in the group of the "more feminine" male music majors). There is also an issue with letting non-experts decide whether or not the information is diagnostic, based only on intuition, as it is not known then if the information is *actually* diagnostic or not.

This issue is re-occurring in a follow-up experiment presented in the same paper (Nisbett et al., 1981). In the follow-up study the pre-test group, whose evaluation of information as clinical or non-clinical was used to decide if information was indeed clinical or not, were master students

in social work. The task at hand was to evaluate if 150 different characteristics were predictive (diagnostic) of child abuse, counterpredictive (meaning a sign that someone is *not* a child abuser, referred to as "counterdiagnostic" in the study) or nondiagnostic (Nisbett et al., 1981). Again the question should be raised if these students actually knew what behaviour/characteristics predicts child abuse or not. The argument could be made that even though they are advanced students in social work they might not actually know child abuse indicators top of mind, but might instead use their intuitive sense of these predictors (or counterpredictors). This is problematic because information that intuitively *seems diagnostic* may in fact not be diagnostic at all (Dana et al., 2013; Tversky & Kahneman, 1973).

The year after Nisbett et al.'s study one of its authors, Zukier (1982), published another study on dilution. In this study the task was to estimate GPA (grade point average) of described students. The control group received only diagnostic information perceived by pretest students to be highly correlated with GPA. This could be hours per week studying, how much the student expected to earn in 15 years time or if they were good at finishing what they've started. Part from this information the participants in the experimental (nondiagnostic) condition also received additional information about the subject to be evaluated that described the subject as "average" on dimensions relating either highly or poorly (meaning, essentially, unrelated) to GPA (again: perceived so by pre-test students). This could be information like how many houseplants or siblings they had (unrelated) or personality ("he once in a while feels like being on his own and being his own boss", related to GPA but average). They found that the additional information diluted the judgment/predictions and that it did so just as much when the added "average"-information correlated to GPA as when it didn't. They concluded that average information, whether relating to the outcome or not is treated as a sign of averageness of the target, reducing the similarity to any extreme performer (Zukier, 1982)

A third famous study on the dilution effect was conducted by Tetlock and Boettger (1989). In one version of their study the participants received either diagnostic information only (amount of hours studied per week) or coupled with nondiagnostic information (of the type "Robert is widely regarded by his friends as being honest," "Robert plays tennis or racquetball about three or four times a month," "Robert describes himself as a cheerful person," and "Two months is the longest period of time Robert has dated one person"). The participants were also told that if they felt they had no useful information, they should simply predict the average GPA for the university

where the experiment took place, approximately 3.0. The results showed that the participants receiving added nondiagnostic information gave predictions significantly closer to the mean (3.0) than in the undiluted condition (receiving only information on how much time the student studied each week). In this study they we're also interested in measuring if accountability (meaning pressure to justify one's views to others) magnified the dilution effect, because the participants would then be forced to put more consideration into their answers and would be more prone to dwell on their judgments. They found that relative to unaccountable subjects the subjects who were made accountable diluted their predictions in response to nondiagnostic information and that they were more responsive to additional diagnostic information. The accountability manipulation made subjects use a wider range of information in making judgments, but it didn't make them better at judging the usefulness of that information (Tetlock & Boettger, 1989). However, this study also treated the information as diagnostic or not based on how it was labelled by the undergraduate students, just as in the previous studies by Nisbett et al. (1981) and Zukier (1982), This means that the same arguments about the reliability of the undergraduate student's knowledge of what information is indeed diagnostic or not can be made here. We do not know for sure if hours studied per week is indeed related to performance, only that it was perceived as strongly related by other students. Albeit more recent studies on the success of people with a high degree of self-control and/or grit, which should be closely related to hours studied per week, might indicate that it indeed is (Duckworth & Gross, 2014). The idea of this method is that as the information was deemed diagnostic or not by the same group (undergraduates) that was later used as subjects (though not the same individuals), so they are expected to agree with themselves. This follows a logic of course (if the students don't know that a piece of information is diagnostic how are they expected to react to it as such?) but the use of the term diagnostic is somewhat different here than how it is usually defined, which is information of high value to reach conclusions about cause and effect.

The issues of diagnosticity in both Nisbett et al.'s (1981) and Tetlock and Boettger's (1989) studies was addressed in a series of studies by Dana et al. (2013) where the dilution effect was measured against actual outcomes in performance. In their first study, either just prior GPA and biographical information, or also a conducted interview was used to predict future GPA. The results showed that the group also conducting interviews did a significantly worse job predicting GPA (r = .31) than the group only receiving biographical information and prior GPA (r = .65). A follow-up study also found that even when the conducted interview generated random answers the

participants still felt that they gained useful information from it. The researchers attribute this to sensemaking. The fact that people are so eager to build coherent stories that they can form these out of basically anything (Dana et al., 2013).

Failure to replicate dilution theory

In 2021 Highhouse et al. did a large study (n = 796) where they failed to replicate the dilution effect. The study was a systematic replication, based on the studies by Nisbett (1981), Tetlock & Boettger (1989) and Zukier (1982) mentioned above. The participants were recruited via Amazon's Mechanical Turk (MTurk), so unlike the previous studies this study used paid participants instead of students. The experiment had four conditions in a 2x2 design and the participants were randomly assigned to one of them. The task was to predict a student's GPA based on the information received. The participants either received only diagnostic information or also three pieces of nondiagnostic information. The diagnostic information was either indicative of a high GPA (hours studied per week = 31) or low GPA (3 hours), the same as used by Zukier (1982) and Tetlock and Boettger (1989). The nondiagnostic information consisted of 12 items where three items we're randomly presented to each participant, see Table 1.

Table 1

Nondiagnostic items, from Highhouse et al. (2021)

Nondiagnostic item	Source
• Has four plants in the place he is living in now • Is the third child in his family • Goes to about four movies during the semester • Spent 6 days camping and hiking last year	Zukier, H. (1982). The dilution effect: The role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality and Social Psychology, 43(6),* 1163–1174.
• Plays tennis or racquetball three or four times a month • Is widely regarded by his friends as being honest • Describes himself as a cheerful person • Two months is the longest period of time (target) has dated one person	Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology, 57(3),* 388–398.
• Spends his summers landscaping for extra money • Is not a big fan of electronic music • Painting is one of (target's) favorite hobbies • Enjoys watching TV in his free time	Highhouse, S., Freier, L. M., Stevenor, B. A., Shea,M. A., Childers,M., & Melick, S. R. (2021). Failure to replicate the basic dilution effect in performance prediction. *International Journal of Selection and Assessment, 1–7.*

They also tested level of cognitive reflection by using the three item cognitive reflection test (Frederick, 2005), with questions like “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?”, which relates to the theory originally formulated by Tetlock and Boettger (1989) that accountability magnifies the dilution effect. The expectation being that participants with a higher degree of reflection would be more prone to overthink their judgments and use the nondiagnostic information in doing so. Part from the prediction the participants were also asked to state how certain they were in their prediction, using a slider from 0 to 100. The results showed no main effect for the presence of nondiagnostic information, nor an interaction effect of the direction of diagnostic information (hours studied) and the presence of nondiagnostic information. There was also no effect when taking into account the participants level of cognitive reflection (as was indicated by Tetlock & Boettger’s, 1989, study). They also found no effect of the presence of nondiagnostic information on confidence in predictions. They did, however, find an effect of the direction of diagnostic evidence (31 hours or

3 hours per week of studying) where participants receiving positive indications (31 hours per week) were more confident than those receiving negative indications (3 hours per week).

The theory of dilution doesn't necessarily concern accuracy and are not to be confused with validity. In the previously mentioned study by Dana et al. (2013) the predictions are compared to an outcome, but this is not the case in the other mentioned studies. Highhouse et al. (2021) also points this out:

> It is important to note that the original dilution effect concerns the tendency for nondiagnostic information to dilute the extremity of predictions that people make when presented only with diagnostic information. This is not the same thing as examining the predictive validity of the GPA predictions. Dilution may or may not impact prediction accuracy, which deals with the rank-order consistency of predicted and achieved outcomes. (p. 2).

Dilution theory hasn't been that widely researched in later years, as Highhouse et al. (2021) also notes in their replication study. Perhaps, they stipulate, due to unpublished null findings. With Highhouse et al.'s recent failure to find any dilution effect in a large (n = 796) sample there is an apparent need to further investigate the effect of nondiagnostic information in the process of evaluative judgment.

Confidence in predictions

Confidence and accuracy are often poorly related in interpersonal prediction contexts (Dunning et al., 1990; Swann & Gill, 1997) and confidence has been shown to increase with information even in situations where accuracy does not (Andersson et al., 2005; Hall et al., 2007). Furthermore, the very fact that someone is incompetent might lead to blindness of one's own shortcomings, resulting in a severe overconfidence, the so-called Dunning-Kruger effect (Kruger & Dunning, 1999). In several of the mentioned studies on dilution the participants are also asked to state their confidence in their rating/evaluation (Dana et al., 2013; Highhouse et al., 2021; Tetlock & Boettger; 1989). The results on this variable have shifted between studies. While Tetlock and Boettger (1989) found that more information reduced confidence, Dana et al. (2013) found that level of confidence was the same or slightly higher if the participants watched or conducted an interview with random answers as an interview with accurate ones. They attribute this to people's habit of sensemaking, the need and ability to create a coherent story out of almost

anything. Lastly, Highhouse et al. (2021) found that confidence was not affected by nondiagnostic information, only by the direction of evidence, where indications of high GPA yielded higher certainty than low GPA indications. The authors stipulated that this might be because the participants thought that many hours studied are bound to result in high grades, while low amount of hours studied might be compensated by other, unknown, individual factors (ease of learning or other).

Research aim and hypotheses

This study aims to further investigate the effect nondiagnostic information has on evaluative judgment. The evaluative task will be to assess perceived job fit for a position the fictitious candidates are applying for, and not, as in previous studies, to predict GPA.

Based on the failure of Highhouse et al. (2021) to replicate the basic dilution effect, this study will investigate if the effect depends on the type of nondiagnostic information added. This is tested by dividing the nondiagnostic information about the evaluative targets that participating judges receive into two categories: obviously nondiagnostic information (referred to as nondiagnostic information) and seemingly diagnostic information (referred to as faux-diagnostic information). This is then compared to judges only receiving the diagnostic information (the diagnostic only group, who serves as control). This study is, however, not primarily interested in the dilution effect per se, as dilution only refers to diminished extremity in judgment. Rather the aim is to see if the added information creates more noise, i.e. variability, in judgment. The reason for this is because the consequence of added noise is larger than if extremity in judgment is affected in a predictable (meaning shared) way across judges. The study will also test if the level of confidence in one's judgment is affected by the type of information received about the candidates.

Hypothesis 1

The type of added information (nondiagnostic or faux-diagnostic) shared about the candidates will have an effect on candidate ranking.

Hypothesis 2

There will be a difference in level of noise depending on what type of information is shared about the candidates (diagnostic only, diagnostic and nondiagnostic or diagnostic and faux-diagnostic).

Hypothesis 3

The type of added information (nondiagnostic or faux-diagnostic) shared about the candidates will have an effect on confidence in judgment.

Method

Participants

The participants were recruited via Prolific (https://www.prolific.co/), an online scientific platform for recruiting research participants. Each participant was compensated with £0.8 for their participation, and the study took in average 9,5 minutes (*SD* = 5.5) to complete. Out of the 151 participants who completed the experiment six extreme outliers (Q3 + 1.5xIQR or Q1 - 1.5xIQR) were excluded. These participants are believed to have only randomly evaluated the candidates in order to receive the compensation. Another indication that something was amiss with some of these participants was that three of them filled out the believed purpose of the study (the only free text item in the study) in very poor English (e.g. “for now how really are you”) and one filled it out in a foreign language, leading to the suspicion of insufficient language skills (the test was in English). The final study consisted of n = 145 participants, 48 in the DO-group, 50 in the FD-groups and 47 in the ND-groups.

The majority of the participants were students (74%) age 18-24 (78%), 50% male, 48% female and 2% identified as non-binary/third gender. The distribution of both age, occupation and gender was evenly dispersed between conditions. The minimum age to participate was 18, and the study recruited only Prolific-users registered as students, however some of them reported their main occupation as employed (24%), or unemployed (2%), at the time of participating in the study.

Material

Depending on which condition they were assigned to the participants received different kind of information about the candidates. The DO-group received only diagnostic information, the

FD-group received the same diagnostic info as the DO-group but also received additional faux-diagnostic (meaning seemingly diagnostic) information, and the ND-group received the same diagnostic information as the other two groups and (apparently) nondiagnostic information about the candidates.

Type of information shared in each condition

Diagnostic information. GMA/IQ, level of conscientiousness and score on situational test was used as diagnostic information.

Faux-Diagnostic Information. In this study a new construct is created that evaluates the nondiagnostic information in two categories: Faux-diagnostic and Nondiagnostic information. The Faux-diagnostic information is information that is in fact not a predictor of job performance, but that can easily be perceived as such. It is also the type of information that is often shared about candidates in a hiring situation. In this study the personality traits Openness to experience, Extroversion and Agreeableness was used as Faux-diagnostic info together with a short self-description from the candidates.

Nondiagnostic Information. For this study the nondiagnostic information was supposed to be apparently irrelevant for the candidates expected job performance. The type of information shared varied between candidates, but care was taken to not reveal any information that could be of any seeming relevance for the position as accountant, and information that triggered bias was avoided (no information shared about gender, age, education or such). The type of information shared was where the candidate grew up (small town, rural are, city), how many siblings he or she has, hobbies, likes and dislikes, favorite food. Some examples of the presentations: “Candidate L enjoys watching old movies and their favorite food is Pizza. They have a dog.”, “Candidate H enjoys shopping and eating out. They have an older sibling and the two of them grew up in a small town.”. The type of information shared was loosely based on non-relevant information shared in studies measuring dilution-effect in selection (Highhouse et al., 2021; Tetlock et al., 1981; Zukier, 1982). See Appendix A for a table of all candidates as described in the study’s three conditions, to get a visual presentation of the complete study please see https://bit.ly/judgcand.

Procedure

The experiment was created using Qualtrics (https://www.qualtrics.com/), an online survey platform. Each participant we’re randomly assigned to one of six groups, two groups for each of

the three different conditions: diagnostic only (group DO1 and DO2), Faux-diagnostic (FD1 and FD2) or Nondiagnostic (ND1 and ND2). The participants we're not informed of the randomization or the other possible conditions of the study.

The participants were asked to assess 15 candidates applying for the position as accountant, and was instructed to take on the role of hiring manager. After receiving information on the nature of the study and ethical information and consenting to the conditions of the study the participants received some information on the instruments used to describe the candidates they we're asked to assess. To make sure they had read and understood the information two control questions we're asked at the end and if the participants got them wrong they got an error message and was asked to try again. After this step they we're asked to fill out demographic information on age, gender and occupation and if they had previous hiring experience.

Each participants rated the same 15 candidates, in random order. After receiving some information about the candidate, the task was stated the same way for each candidate and in each condition "Based on the information above, please assess the candidate's fit for the position and state how certain you are in your assessment.". Both assessment and level of confidence was measured on an integer scale ranging from 0-100, with 5 as integer (possible values: 0, 5, 10, 15 and so on).

At the end of the study the participants got to state which piece of information they found most useful for assessing the candidates and what they perceived the purpose of the study to be. At the very last step they we're thanked for their participation and received contact information (again, this was also shared at the beginning) if they had any questions or comments for me as a researcher.

Statistical analysis

The statistical analysis was performed in R Studio (R Core Team, 2021). The figures were produced using the package ggplot2 (Wickham, 2016).

As the two groups receiving diagnostic info only (DO1 and DO2) had no individual difference and for all purposes of the analysis can be considered the same condition they were later combined into one group (the order of the information presented varied between the groups and within the group, thus evening out any order-effect).

A one-way ANOVA tested whether there were any order effects of information (FD1 and FD2 received the information in different order, as well as ND1 and ND2). As no significant differences were found between the two groups within each condition for either ranking correlation, FD: $t(48) = -0.63$, $p = .53$, ND: $t(45) = -1.22$, $p = .23$ or mean confidence levels, FD: $t(48) = 0.39$, $p = .70$, ND: $t(45) = -0.94$, $p = .35$ the groups ND1 and ND2, and FD1 and FD2 were combined to gain statistical power. All further analysis was carried out with only three groups, one group per condition (DO, FD and ND). Within each condition the group number (1 or 2) decided in which order the information was received. For an overview of information shared with each group see Table 2.

Table 2

Type of Information and Order of Information Shared with Each Group

Group	Type of information (in order)
DO (DO1 & DO2)	Diagnostic - IQ, situational test, conscientiousness
FD 1	Personality traits - openness to experience, conscientiousness, extroversion, agreeableness Diagnostic – IQ, situational test Self-description
FD 2	Diagnostic – IQ, situational test Self-description Personality traits - openness to experience, conscientiousness, extroversion, agreeableness
ND 1	Diagnostic - IQ, situational test, conscientiousness Nondiagnostic – vignette about hobbies, favorite food, siblings and/or similar
ND 2	Nondiagnostic – vignette about hobbies, favorite food, siblings and/or similar Diagnostic - IQ, situational test, conscientiousness

Notes: DO = Diagnostic Only (Group), FD = Faux-Diagnostic (Group), ND = NonDiagnostic (Group)

For a full list of how each candidate was presented in each group and condition see Appendix A.

To compare the ranking of the candidates in the different conditions and to analyse the data, a "base ranking" for the candidate's was needed. As the study is not intended to evaluate the

judgments (stating which group does "better" predictions) and have no way of claiming an "ideal" ranking of the candidates (we have no known outcome of the candidates later performance), the control group's (the DO-group) ranking was used as base. The base ranking sets the mean top candidate in the DO-group as number one, the second as number two and so forth. Each participant in the study then got a correlation score of how well their ranking corresponded with the base ranking, and these correlation-scores was then compared between conditions.

A reason for this, to use between-candidate *ranking* instead of *ratings*, is to keep the relevance high, as the effect of a between-judge difference in ranking (meaning the ordering of candidates from best suited to worst) is higher than a difference in rating (absolute numerical scores). This is perhaps best illustrated with a real-world example: If only three candidates are selected for the next step in the recruiting process, it matters more which three candidates are scored highest (and thus selected) than which numerical score they receive. If there is no difference between the conditions then the same three candidates would be ranked highest, in essence selected for (e.g.) further interviews, no matter what kind of information about them was shared with the recruiter in the initial screening process. There would be no effect on the outcome of the screening process. If the opposite is true then different conditions would yield a different set of candidates. Conclusively, which candidates are selected is more important than if the rating/scores of the candidates differ, it doesn't matter for the outcome if a certain candidate gets a score of 87 or 95, but how they are ranked compared to the other candidates.

Ethical Considerations

This study adopted and followed the ethical rules of the Swedish Ethical Review Authority. The participants were hence informed that the study did not 1) collect sensitive personal data (such as political views) that could be tied to a specific person, 2) use a method that included physical intrusion, 3) use a method that aimed to affect the participants physically or psychologically, 4) implicate a risk for psychological or physical damage. Participants were also informed about their right to withdraw at any time. This information was shared in a consent form before starting the study. The participants also received contact information to both the researcher and the assigned supervisor.

Results

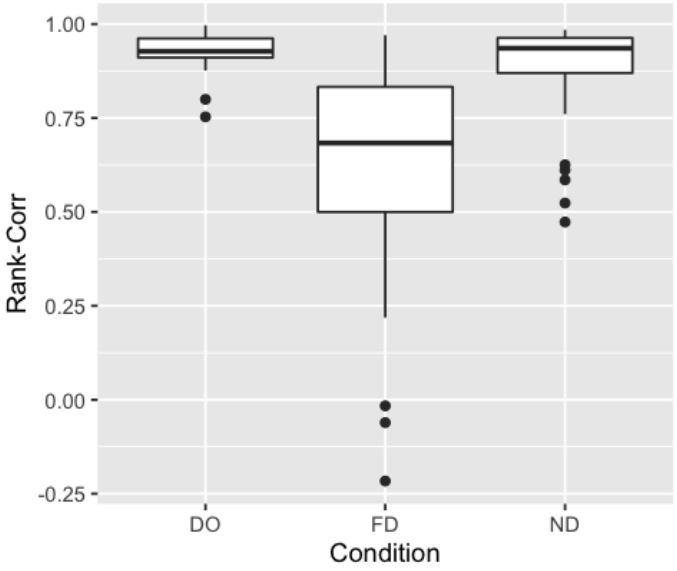
Hypothesis 1

The first hypothesis concerns the effect of type of added information on ranking. To test this a one-way ANOVA was done, comparing the base ranking-correlation between conditions (as described under *statistical analysis* above). As the data was not normally distributed (a Shapiro Wilks test returned significant p-values in all three conditions) a Kruskal-Wallis test was performed in order to add credibility to the significance test.

The ANOVA showed that the groups differed significantly, $F(2, 142) = 39.89$, $p < .001$, eta-squared 0.36, which indicates support for the first hypothesis of the study. There was a significant effect of the added information in the FD-group ($M = 0.63$, $SD = 0.28$) but not in the ND-group ($M = 0.89$, $SD = 0.13$), the added obviously irrelevant information shared in the ND group had no significant effect ($p = .24$) on how the participants were judged (in comparison to each other, so no difference in ranking), while the seemingly relevant information did have an effect ($p < .001$). This result is visualized in a boxplot in Figure 1, where the added variation in the group receiving faux-diagnostic information is also apparent. For a more detailed overview of the ratings and standard deviations per candidate and condition see Appendix B.

Figure 1

Boxplot of Ranking Correlation per Condition



Notes: DO = Diagnostic Only (Group), FD = Faux-Diagnostic (Group), ND = NonDiagnostic (Group)

The Kruskal-Wallis test showed that the differences were indeed significant, X^2(2, N = 145) = 57.75, p < .001. A post-hoc Dunn test confirmed that the difference was significant only between the FD group and the other two groups, FD – DO Z = 7.04, p < .001, FD – ND Z = -5.97, p < .001, and not between the DO and ND group, Z = 1.02, p = .15. Neither age, previous hiring experience nor gender had any effect on the ranking correlation in any condition.

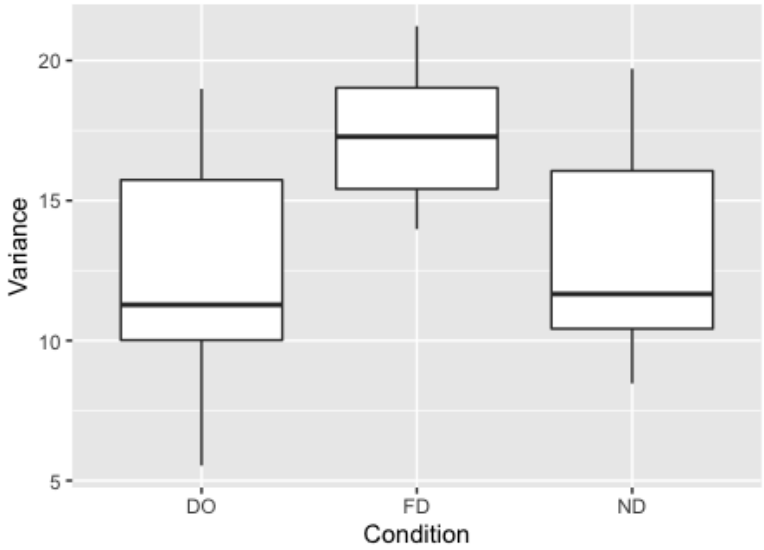
The data was also checked for how often the three strongest candidates (candidate M, B and N) was rated highest. This mirrors a real-world situation where the three strongest candidates are selected for the next step in the recruitment process. In the DO-group the three top candidates were also ranked as top 3 in 69% of cases, in the ND-group they were included in 62% of the cases, in the FD-group only 10% of judges ranked M, N and B as top three.

Hypothesis 2

To test the second hypothesis, the difference in amount of noise depending on type of added information, a one-way ANOVA was conducted that compared the mean standard deviations in ratings of each candidate between different conditions (see Appendix B). The amount of variance in judgment was significantly higher in the FD condition (M = 17.5) than in the ND (M = 13) or DO (M = 12.45) conditions, F(2, 48) = 9.49, p < .001, eta-squared = 0.31. The assumptions of normality and homogeneity in the sample were met. The amount of variance is illustrated in a boxplot, see Figure 2.

Figure 2

Boxplot of Variance per Condition



Notes: DO = Diagnostic Only (Group), FD = Faux-Diagnostic (Group), ND = NonDiagnostic (Group)

However, when comparing different groups of candidates (top, medium and bottom rated) the amount of variance was only significantly larger in the group with the five top candidates (candidates M, N, B, C and L), $F(2, 12) = 45.82$, $p < .001$, eta-squared = 0.88.

Hypothesis 3

Hypothesis 3 stated that type of added information would affect the level of confidence in judgment. This hypothesis was tested by analysing the variance in mean confidence levels of participants in the different conditions. The assumptions for normality and homogeneity in the sample were met. A one-way ANOVA showed that there was a significant effect of type of information shared on confidence, $F(2, 142) = 8.20$, $p < .001$, eta-squared = .10. The highest confidence was from participants in the DO-group ($M = 80.47$, $SD = 11.33$), the second highest in the ND group ($M = 75.12$, $SD = 15.10$) and the lowest level of confidence was found in the FD group ($M = 68.67$, $SD = 16.36$). There was no effect of age, gender or previous hiring experience on level of certainty.

Discussion

Hypothesis 1

Hypothesis 1 stated that the ranking of candidates would be affected by the type of added information, and this was also supported in the result. The nondiagnostic information didn't significantly affect the ranking, while the faux-diagnostic information did. The fact that the three strongest candidates was only selected in 10% of cases in the FD-group also shows that the seemingly diagnostic information has a large effect on the outcome in an actual recruiting situation, where the top candidates are selected for the next step in the recruiting process. This implicates that seemingly diagnostic information could play a big part in who gets selected in an initial screening of candidates when the evaluator has access to self-descriptive data (as usually shared in a cover letter) or personality traits not related to performance.

It might also be that the added information about the otherwise strongest candidates, that showed them as average in other ares created the sense of "averageness" that Zukier (1982) claims contributes to the dilution effect. Meaning that the similarity between the candidate and a top performer was reduced, as the candidate proved to be average on other areas, even though these areas are not related to performance, following the similarity model by Tversky (1977). In Nisbett et al's (1981) study they stipulate that with the nondiagnostic information the person described suddenly is less similar to the (culturally shared) prototype, in this case the prototype of an accountant.

The fact that there was no difference between the nondiagnostic (ND-group) and the diagnostic only (DO) group also mirrors the findings from Highhouse et al. (2021) who failed to replicate the dilution effect (or rather, who found no effect whatsoever, dilution or other) when the added information was of the kind "Is the third child in his family" or "Enjoys watching TV in his free time", as represented in this study by the added information shared with the ND-group. The finding in this study and the findings of Highhouse et al. (2021) indicates that more studies are needed on when, how and what kind of nondiagnostic information affects judgment.

Hypothesis 2

The amount of noise, or variance, was much larger in the group also receiving faux-diagnostic information than in the group also receiving nondiagnostic information or the group receiving only diagnostic information, which gives support to our second hypothesis, that type of

added information affects noise. The findings indicate that apparent nondiagnostic information doesn't add noise (and thus is, in that sense, "harmless") while seemingly diagnostic information does. This is reasonably because the added information influence judges differently. Some may favour extroversion highly, perceiving it to be of great relevance to the performance of an accountant, while others don't consider it to be relevant at all. This creates a greater variability in judgments, or added noise, which is always unwanted. The reason that noise is always unwanted is simple, as a noisy process is unpredictable, and the effect of the individual judge gets bigger than can be reasonably accepted (Kahneman et al., 2021). If you apply for a job, you do not want the reason for being selected or not to be that you got the right (or wrong) evaluator, you want the selection process to be foreseeable and transparent. The same, of course, goes for the employer, who wants the objectively best candidate, not the subjectively best.

Hypothesis 3

We found an effect of non-diagnostic information on confidence, where added information reduced confidence, especially in the group receiving faux-diagnostic information. This mirrors the findings in Tetlock and Boettgers (1989) study, where level of confidence was reduced by added information uncorrelated to predicted outcome (nondiagnostic information). This result is in line with the basic concept of dilution as loss of similarity. The added information described the candidates as average or low scoring on at least one personality trait (see Appendix A for an overview of each candidate), which might reduce the similarity between either an ideal or really poor fit employee. Hence the uncertainty goes up, because you can no longer hold on to the idea that this candidate is either entirely flawless or all bad, they become more nuanced. It is positive that the added irrelevant information didn't create a false sense of certainty in the judges, even if it should, logically, be the same in all three conditions.

Limitations

One limitation to the generalizability of the findings of this study is that it used student participants and not experts, which in this case would be professional recruiters. This was the original idea of this study but finding a satisfactory number of participating professional recruiters wasn't feasible under the time and budget constraints. We do not know how much of the effect on

either ranking, noise or confidence remains if the judge is a professional recruiter, and thus we do not know the implications or harm of seemingly diagnostic information in those situations.

Another limitation is that the setting doesn't mirror a hiring situation very well in how the information is presented to the evaluative judge. Future studies could closer mimic how the information is presented to the judge in an actual hiring situation, for example by using cover letters.

This experiment is based on predictive judgment with no known outcome, as all candidates are fictious, and it would have been valuable to know the difference in actual quality of predictions, as a suggestion by using past recruiting data. This would enable the use of a measurement (post-hiring evaluation by manager or time spent in position or such) as an outcome measure of the actual success or fit of the candidate. This was also something that was thought of for this study, but the request to use data about previous hires, made to a recruiting firm that was contacted for possible collaboration, was denied with reference to jurisdiction on the use of personal information (GDPR).

There are also some important differences between this experiment and the previous ones by Nisbett et al. (1981), Tettlock & Boettger (1989), Zukier (1982) and Highhouse et al. (2021) in that the participants in this study were judging the candidates not on predicted GPA but on perceived job fit. This could have affected the outcome, as an example perhaps the performance of a student is perceived to more likely be influenced by personal (nondiagnostic) factors than the performance of an accountant? Meaning the participants are rightfully more swayed by nondiagnostic info in that context, however that doesn't explain why Highhouse et al. (2021) failed to replicate the dilution effect with apparently nondiagnostic information, as they also measured predicted GPA. It can also simply be easier to predict GPA than job performance due to a reduced number of factors influencing (or seemingly influencing) the outcome (Dana et al, 2013). Another difference is that the number of items to score were much larger in this study (15) than in Nisbett et al. (2), Tettlock and Boettger (1), Highhouse et al. (1) and Zukier (6). This might affect how much thought and consideration are put into each evaluation. If asked to make many consecutive judgments you might use a different method/heuristic to quickly evaluate each candidate (zeroing in on the information you perceive most useful) than if you are only making one or a few judgments. If asked to rate only one or a few cases you might instead be more inclined to think hard about your decision, which in itself is believed to lead to increased dilution (Tettlock &

Boettger, 1989). However the increased amount of measures per participant could also have increased the reliability of the study and added power to it's findings. But because of these differences caution should be taken when comparing the results from this study with previous dilution-studies.

Suggestions for further research

A recommendation for future research is to use professionals instead of students, as discussed in the limitations-section above. As many recruiting decisions are not made by professional recruiters but by managers, both groups should be used as subjects to evaluate the effect of seemingly diagnostic information in hiring. Another suggestion for future research, dealing with the issue of unknown outcome, would be to use empirical samples of past recruiting, as also described in the limitations-section above. Third it would be interesting to isolate the effect different type of faux-diagnostic information has on judgment. Perhaps the self-descriptions influenced the participants more, or a certain personality measure, like extroversion, is a heavier anchor than others. In other words to see if the participants are biased. Fourth, it would be interesting to further investigate the effect of nondiagnostic information, it could be that nondiagnostic information has a dilution effect if it is presented verbally by the subject themselves, as is the case in an interview, or that if you blend it with the diagnostic information it is harder to cognitively disregard it.

Conclusion

This study indicates that while apparently nondiagnostic information doesn't affect judgment, seemingly diagnostic information does. Even though we can't draw any conclusions about the accuracy of predictions in the different conditions, as we do not have an outcome, we can conclude that the type of information shared about candidates' matter, and that it affects how we judge them. We can also conclude that added faux-diagnostic information creates more noise (and noise is always unwanted). As confirmed in previous studies personality traits paired with job analysis can be predictive of performance (Tett & Rothstein, 1991), but in general, when personality tests are used, the candidates are arbitrarily tested on all the big 5 personality traits (Moy & Lam, 2004). This study indicates that if some personality traits are unrelated to performance, a candidate evaluation shouldn't include results on those traits, as it creates noise.

This is because different judges will value different traits to be of higher or lower relevance, leading to a higher variance in ratings. Another reason is that the evidence at hand no longer points in one single direction (the candidate is all of a sudden strong in some areas while weak in others). Multiple conflicting cues creates ambiguity, and ambiguity is the reason that complex problems (with conflicting cues, such as cues about both excellence and averageness) is noisier than simple ones (Kahneman et al., 2021). Low inter-rater reliability, when systematic, is referred to by Kahneman et al. (2021) as "system noise" and they explain the problem with this in an elegant way "System noise is inconsistency, and inconsistency damages the credibility of the system" (p. 53). Damaged credibility in the area of recruiting creates uncertainty, for the hiring party (did we get the best candidate or was it chosen based on irrelevant factors?) as well as for the candidate's (did I not get the job due to lack of ability, or did I just get the wrong evaluator?). To reduce noise in hiring, benefitting both candidates and hiring organizations, a stricter conduct in the selection process is recommended.

References

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology, 96(4)*, 790-805.

Andersson, P., Edman, J., & Ekman, M. (2005). Predicting the world cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting, 21*, 565–576.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology, 44*(1), 1-26.

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology, 94*(6), 1394-1411.

Boston Consulting Group (2007). The future of HR: Key challenges through 2015. BCG.

Breaugh, J. A., & Starke, M. (2000). Research on employee recruitment: So many studies, so many remaining questions. *Journal of management, 26*(3), 405-434

Brunswik, E. (1952) The conceptual framework of psychology. *International encyclopedia of unified science.* University of Chicago Press.

Campion, M. C., Ployhart, R. E., & MacKenzie Jr, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*(4), 283-310.

Chambers, E. G., Foulon, M., Handfield-Jones, H., Hankin, S. M., & Michaels III, E. G. (1998). The war for talent. *The McKinsey Quarterly*, (3), 44-57.

Chan, D., & Schmitt, N. (2002). Situational Judgment and Job Performance. *Human Performance*, *15*(3), 233-254.

Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*(5), 512–587.

Dalal, D. K., Sassaman, L & Zhu, X. (2020) The Impact of Nondiagnostic Information on Selection Decision Making: A Cautionary Note and Mitigation Strategies, *Personnel Assessment and Decisions, 6/2*(7), 54-64

Duckworth, A., & Gross, J. J., 2014, Self-Control and Grit: Related but Separable Determinants of Success. *Current Directions in Psychological Science, 23(5*), 319-325.

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of personality and social psychology*, *58*(4), 568-581.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Gorman, E. H. (2005). Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms. *American Sociological Review*, *70*(4), 702-728.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology*, *18*, 135– 164.

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, *103*, 277–290.

He, Y., Donnellan, M. B., & Mendoza, A. M. (2019). Five-factor personality domains and job performance: A second order meta-analysis. *Journal of Research in Personality*, *82*, Article 103848.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333-342.

Highhouse, S., Freier, L. M., Stevenor, B. A., Shea, M. A., Childers, M., & Melick, S. R. (2021). Failure to replicate the basic dilution effect in performance prediction. *International Journal of Selection and Assessment 30*(2), 195-201.

Johnson, J. A. (1981). The" self-disclosure" and" self-presentation" views of item response dynamics and personality scale validity. *Journal of Personality and Social Psychology*, *40*(4), 761-769.

Kausel, E. E., Culbertson, S. S., & Madrid, H. P. (2016). Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring decisions. *Organizational Behavior and Human Decision Processes*, *137*, 27-44.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121-1134.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: a meta-analysis. *Journal of applied psychology*, *98*(6), 1060-1072.

Kuncel, N. R., Klieger, D. M., & Ones, D. S. (2014). In hiring, algorithms beat instinct. *Harvard business review*, *92(5)*, 32-32.

Levy, J. J., Richardson, J. D., Lounsbury, J. W., Stewart, D., Gibson, L. W., & Drost, A. W. (2011). Personality Traits and Career Satisfaction of Accounting Professionals. *Individual Differences Research*, *9*(4), 238-249.

MacKenzie, W. I., Ployhart, R. E., Weekley, J. A., & Ehlers, C. (2009). Contextual Effects on SJT Responses: An Examination of Construct Validity and Mean Differences Across Applicant and Incumbent Contexts. *Human Performance*, *23*(1), 1-21.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of applied psychology*, *79*(4), 599-616.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, *60*(1), 63-91.

Momin, W. Y. M., & Mishra, K. (2015). HR analytics as a strategic workforce planning. *International Journal of Applied Research*, *1*(4), 258-260.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel psychology*, *60*(3), 683-729.

Moy, J.W., Lam, K.F. (2004), Selection criteria and the impact of personality on getting hired, *Personnel Review*, *33*(5), 521-535.

Neckerman, K. M., & Kirschenman, J. (1991). Hiring strategies, racial bias, and inner-city workers. *Social problems*, *38*(4), 433-447.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive psychology*, *13*(2), 248-277.

Onwuegbuzie, A. J., & Daley, C. E. (2001). Racial Differences in IQ Revisited: A Synthesis of Nearly a Century of Research. *Journal of Black Psychology*, *27*(2), 209–220.

Perlow, R. & Kopp, L. (2004) Conscientiousness and Ability as Predictors of Accounting Learning, *Human Performance, 17:4*, 359-373.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g.. *Journal of applied psychology*, *79*(4), 518-524.

Rivera, L. A. (2012). Hiring as cultural matching: The case of elite professional service firms. *American sociological review*, *77(6)*, 999-1022.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human resource management review*, *16*(2), 155-180.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, *124*(2), 262-274.

Schmidt, F. L., & Hunter, J. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology*, *86*(1), 162–173.

Schmidt, F. L. & Oh, I-S. (2021). Select on Intelligence. *Fox School of Business Research Paper* (in press), http://dx.doi.org/10.2139/ssrn.3961855

SHRM (2017). *SHRM Customized Talent Acquisition Benchmarking Report. All industries, all sizes.* Available for download (2021-03-28): http://shrm.org/

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and Reliability of Situational Judgement Test Scores: A New Approach Based on Cognitive Diagnosis Models. *Organizational Research Methods*, *19*(3), 506–532.

Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The Predictive Value of IQ. *Merrill-Palmer Quarterly*, *47*(1), 1–41.

Swann, W. B., Jr., & Gill, M. J. (1997). Confidence and accuracy in person perception: Do we know what we think we know about our relationship partners? *Journal of Personality and Social Psychology*, *73*(4), 747–757.

Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology, 57*, 388-398.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel psychology*, *60*(4), 967-993.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel psychology*, *44*(4), 703-742.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.

Tversky, A. (1977) Features of similarity. *Psychological Review, 84*, 327-352.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag.

Zukier, H. (1982). The dilution effect: The role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality and Social Psychology*, *43*(6), 1163–1174.

Appendix A: Table of the 15 candidates as presented in each condition.

Candidate M	
DO1	Candidate M took a general aptitude (IQ) test and got a score of 131. The candidate scored 9/10 on conscientiousness. They took a situational test and was assessed as "Excellent".
DO2	Candidate M scored 9/10 on conscientiousness. The candidate took a situational test and was assessed as "Excellent". They took a general aptitude (IQ) test and got a score of 131.
FD1	Candidate M took a personality test and scored: - 6/10 on "openness to experience", - 9/10 on conscientiousness, - 3/10 on extroversion and - 4/10 on agreeableness. Candidate M took a situational test and was assessed as "Excellent". They took a general aptitude (IQ) test and got a score of 131. They describe themselves as ambitious and knowledgeable.
FD2	Candidate M took a situational test and was assessed as "Excellent". They took a general aptitude (IQ) test and got a score of 131. They describe themselves as ambitious and knowledgeable. Candidate M took a personality test and scored: - 6/10 on "openness to experience", - 9/10 on conscientiousness, - 3/10 on extroversion and - 4/10 on agreeableness.
ND1	Candidate M took a general aptitude (IQ) test and got a score of 131. The candidate scored 9/10 on conscientiousness. They took a situational test and was assessed as "Excellent". Candidate M grew up in a household of three. They like Indian food and when it's time to relax they like to watch movies.
ND2	Candidate M grew up in a household of three. They like Indian food and when it's time to relax they like to watch movies. Candidate M scored 9/10 on conscientiousness. The candidate took a situational test and was assessed as "Excellent". They took a general aptitude (IQ) test and got a score of 131.
Candidate N	
DO1	Candidate N scored 8/10 on conscientiousness. The candidate took a situational test and was assessed as "Excellent". They took a general aptitude (IQ) test and got a score of 127.
DO2	Candidate N took a situational test and was assessed as "Excellent". They took a general aptitude (IQ) test and got a score of 127. The candidate scored 8/10 on conscientiousness.

Appendix A

FD1 Candidate N took a personality test and scored:
- 3/10 on "openness to experience",
- 8/10 on conscientiousness,
- 5/10 on extroversion and
- 4/10 on agreeableness.

Candidate N took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 127.
They describe themselves as kind and with a good sense of humour.

FD2 Candidate N took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 127.
They describe themselves as kind and with a good sense of humour.

Candidate N took a personality test and scored:
- 3/10 on "openness to experience",
- 8/10 on conscientiousness,
- 5/10 on extroversion and
- 4/10 on agreeableness.

ND1 Candidate N scored 8/10 on conscientiousness.
The candidate took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 127.

Candidate N grew up in a rural town and now lives in an apartment with their partner. Their favorite sports team is Juventus.

ND2 Candidate N grew up in a rural town and now lives in an apartment with their partner. Their favorite sports team is Juventus.

Candidate N took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 127.
The candidate scored 8/10 on conscientiousness.

Candidate B

DO1 Candidate B took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 122.
The candidate scored 9/10 on conscientiousness.

DO2 Candidate B took a general aptitude (IQ) test and got a score of 122.
The candidate scored 9/10 on conscientiousness.
They took a situational test and was assessed as "Excellent".

FD1 Candidate B took a personality test and scored:
- 7/10 on "openness to experience",
- 9/10 on conscientiousness,
- 7/10 on extroversion and
- 3/10 on agreeableness.

Candidate B took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 122.
They describe themselves as fun loving and cheerful.

FD2 Candidate B took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 122.

They describe themselves as fun loving and cheerful.

Candidate B took a personality test and scored:
- 7/10 on "openness to experience",
- 9/10 on conscientiousness,
- 7/10 on extroversion and
- 3/10 on agreeableness.

ND1 Candidate B took a situational test and was assessed as "Excellent".
They took a general aptitude (IQ) test and got a score of 122.
The candidate scored 9/10 on conscientiousness.

Candidate B enjoys cooking for their friends, preferably Italian. They love dogs and hate stormy weather.

ND2 Candidate B enjoys cooking for their friends, preferably Italian. They love dogs and hate stormy weather.

Candidate B took a general aptitude (IQ) test and got a score of 122.
The candidate scored 9/10 on conscientiousness.
They took a situational test and was assessed as "Excellent".

Candidate C

DO1 Candidate C took a general aptitude (IQ) test and got a score of 120.
The candidate scored 7/10 on conscientiousness.
They took a situational test and was assessed as "Very good".

DO2 Candidate C took a general aptitude (IQ) test and got a score of 120.
The candidate scored 7/10 on conscientiousness.
They took a situational test and was assessed as "Very good".

FD1 Candidate C took a personality test and scored:
- 5/10 on "openness to experience",
- 9/10 on conscientiousness,
- 3/10 on extroversion and
- 5/10 on agreeableness.

Candidate C took a situational test and was assessed as "Very good".
They took a general aptitude (IQ) test and got a score of 120.
They describe themselves as a fast learner who takes their job seriously.

FD2 Candidate C took a situational test and was assessed as "Very good".
They took a general aptitude (IQ) test and got a score of 120.
They describe themselves as a fast learner who takes their job seriously.

Candidate C took a personality test and scored:
- 5/10 on "openness to experience",
- 9/10 on conscientiousness,
- 3/10 on extroversion and
- 5/10 on agreeableness.

ND1 Candidate C took a general aptitude (IQ) test and got a score of 120.
The candidate scored 7/10 on conscientiousness.
They took a situational test and was assessed as "Very good".

Appendix A

	Candidate C grew up in a suburb with one older sibling. They love hiking and their favourite sports team is Real Madrid.
ND2	Candidate C grew up in a suburb with one older sibling. They love hiking and their favourite sports team is Real Madrid. Candidate C took a general aptitude (IQ) test and got a score of 120. The candidate scored 7/10 on conscientiousness. They took a situational test and was assessed as "Very good".
Candidate L	
DO1	Candidate L scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Very good". They took a general aptitude (IQ) test and got a score of 119.
DO2	Candidate L scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Very good". They took a general aptitude (IQ) test and got a score of 119.
FD1	Candidate L took a personality test and scored: - 2/10 on "openness to experience", - 7/10 on conscientiousness, - 5/10 on extroversion and - 7/10 on agreeableness. Candidate L took a situational test and was assessed as "Very good". They took a general aptitude (IQ) test and got a score of 119. They describe themselves as reliable and stress resistant.
FD2	Candidate L took a situational test and was assessed as "Very good". They took a general aptitude (IQ) test and got a score of 119. They describe themselves as reliable and stress resistant. Candidate L took a personality test and scored: - 2/10 on "openness to experience", - 7/10 on conscientiousness, - 5/10 on extroversion and - 7/10 on agreeableness.
ND1	Candidate L scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Very good". They took a general aptitude (IQ) test and got a score of 119. Candidate L enjoys watching old movies and their favourite food is Pizza. They have a dog.
ND2	Candidate L enjoys watching old movies and their favourite food is Pizza. They have a dog. Candidate L scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Very good". They took a general aptitude (IQ) test and got a score of 119.
Candidate J	

Appendix A

DO1	Candidate J took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 106. The candidate scored 8/10 on conscientiousness.
DO2	Candidate J took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 106. The candidate scored 8/10 on conscientiousness.
FD1	Candidate J took a personality test and scored: - 5/10 on "openness to experience", - 8/10 on conscientiousness, - 7/10 on extroversion and - 8/10 on agreeableness. Candidate J took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 106. They describe themselves as someone with high integrity who is detail oriented.
FD2	Candidate J took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 106. They describe themselves as someone with high integrity who is detail oriented. Candidate J took a personality test and scored: - 5/10 on "openness to experience", - 8/10 on conscientiousness, - 7/10 on extroversion and - 8/10 on agreeableness.
ND1	Candidate J took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 106. The candidate scored 8/10 on conscientiousness. Candidate J loves sports. They grew up in the city and their favourite sports team is Juventus.
ND2	Candidate J loves sports. They grew up in the city and their favourite sports team is Barcelona FC. Candidate J took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 106. The candidate scored 8/10 on conscientiousness.

Candidate H	
DO1	Candidate H took a general aptitude (IQ) test and got a score of 105. The candidate scored 6/10 on conscientiousness. They took a situational test and was assessed as "Good".
DO2	Candidate H took a general aptitude (IQ) test and got a score of 105. The candidate scored 6/10 on conscientiousness. They took a situational test and was assessed as "Good".
FD1	Candidate H took a personality test and scored: - 9/10 on "openness to experience", - 6/10 on conscientiousness, - 8/10 on extroversion and - 5/10 on agreeableness.

	Candidate H took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 105. They describe themselves as outgoing and fun.
FD2	Candidate H took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 105. They describe themselves as outgoing and fun. Candidate H took a personality test and scored: - 9/10 on "openness to experience", - 6/10 on conscientiousness, - 8/10 on extroversion and - 5/10 on agreeableness.
ND1	Candidate H took a general aptitude (IQ) test and got a score of 105. The candidate scored 6/10 on conscientiousness. They took a situational test and was assessed as "Good". Candidate H enjoys shopping and eating out. They have an older sibling and the two of them grew up in a small town.
ND2	Candidate H enjoys shopping and eating out. They have an older sibling and the two of them grew up in a small town. Candidate H took a general aptitude (IQ) test and got a score of 105. The candidate scored 6/10 on conscientiousness. They took a situational test and was assessed as "Good".
Candidate G	
DO1	Candidate G scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 101.
DO2	Candidate G scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 101.
FD1	Candidate G took a personality test and scored: - 8/10 on "openness to experience", - 7/10 on conscientiousness, - 4/10 on extroversion and - 5/10 on agreeableness. Candidate G took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 101. They describe themselves as a good co-worker.
FD2	Candidate G took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 101. They describe themselves as a good co-worker. Candidate G took a personality test and scored: - 8/10 on "openness to experience", - 7/10 on conscientiousness,

	- 4/10 on extroversion and - 5/10 on agreeableness.
ND1	Candidate G scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 101.
	Candidate G loves running and hanging out with friends. Their favourite colour is red and they have a cat.
ND2	Candidate G loves running and hanging out with friends. Their favourite colour is red and they have a cat.
	Candidate G scored 7/10 on conscientiousness. The candidate took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 101.
Candidate F	
DO1	Candidate F took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 99. The candidate scored 6/10 on conscientiousness.
DO2	Candidate F took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 99. The candidate scored 6/10 on conscientiousness.
FD1	Candidate F took a personality test and scored: - 3/10 on "openness to experience", - 6/10 on conscientiousness, - 5/10 on extroversion and - 7/10 on agreeableness.
	Candidate F took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 99. They describe themselves as reliable and detail-oriented.
FD2	Candidate F took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 99. They describe themselves as reliable and detail-oriented.
	Candidate F took a personality test and scored: - 3/10 on "openness to experience", - 6/10 on conscientiousness, - 5/10 on extroversion and - 7/10 on agreeableness.
ND1	Candidate F took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 99. The candidate scored 6/10 on conscientiousness.
	Candidate F likes Indian food. In the weekends they love baking and reading.
ND2	Candidate F likes Indian food. In the weekends they love baking and reading.
	Candidate F took a situational test and was assessed as "Good".

	They took a general aptitude (IQ) test and got a score of 99. The candidate scored 6/10 on conscientiousness.
Candidate D	
DO1	Candidate D took a general aptitude (IQ) test and got a score of 97. The candidate scored 5/10 on conscientiousness. They took a situational test and was assessed as "Good".
DO2	Candidate D took a general aptitude (IQ) test and got a score of 97. The candidate scored 5/10 on conscientiousness. They took a situational test and was assessed as "Good".
FD1	Candidate D took a personality test and scored: - 3/10 on "openness to experience", - 5/10 on conscientiousness, - 7/10 on extroversion and - 5/10 on agreeableness. Candidate D took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 97. They describe themselves as prestigeless and people oriented.
FD2	Candidate D took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 97. They describe themselves as prestigeless and people oriented.
	Candidate D took a personality test and scored: - 3/10 on "openness to experience", - 5/10 on conscientiousness, - 7/10 on extroversion and - 5/10 on agreeableness.
ND1	Candidate D took a general aptitude (IQ) test and got a score of 97. The candidate scored 5/10 on conscientiousness. They took a situational test and was assessed as "Good".
	Candidate D is really into yoga. They love cooking and have two dogs.
ND2	Candidate D is really into yoga. They love cooking and have two dogs.
	Candidate D took a general aptitude (IQ) test and got a score of 97. The candidate scored 5/10 on conscientiousness. They took a situational test and was assessed as "Good".
Candidate S	
DO1	Candidate S scored 6/10 on conscientiousness. The candidate took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 88.
DO2	Candidate S scored 6/10 on conscientiousness. The candidate took a situational test and was assessed as "Good". They took a general aptitude (IQ) test and got a score of 88.

Appendix A

FD1 Candidate S took a personality test and scored:
- 6/10 on "openness to experience",
- 6/10 on conscientiousness,
- 5/10 on extroversion and
- 9/10 on agreeableness.

Candidate S took a situational test and was assessed as "Good".
They took a general aptitude (IQ) test and got a score of 88.
They describe themselves as righteous and detailed.

FD2 Candidate S took a situational test and was assessed as "Good".
They took a general aptitude (IQ) test and got a score of 88.
They describe themselves as righteous and detailed.

Candidate S took a personality test and scored:
- 6/10 on "openness to experience",
- 6/10 on conscientiousness,
- 5/10 on extroversion and
- 9/10 on agreeableness.

ND1 Candidate S scored 6/10 on conscientiousness.
The candidate took a situational test and was assessed as "Good".
They took a general aptitude (IQ) test and got a score of 88.

Candidate S grew up in the city with two siblings, their favourite colour is green and they like American food.

ND2 Candidate S grew up in the city with two siblings, their favourite colour is green and they like American food.

Candidate S scored 6/10 on conscientiousness.
The candidate took a situational test and was assessed as "Good".
They took a general aptitude (IQ) test and got a score of 88.

Candidate A

DO1 Candidate A took a situational test and was assessed as "Weak".
They took a general aptitude (IQ) test and got a score of 85.
The candidate scored 5/10 on conscientiousness.

DO2 Candidate A took a situational test and was assessed as "Weak".
They took a general aptitude (IQ) test and got a score of 85.
The candidate scored 5/10 on conscientiousness.

FD1 Candidate A took a personality test and scored:
- 3/10 on "openness to experience",
- 5/10 on conscientiousness,
- 2/10 on extroversion and
- 8/10 on agreeableness.

Candidate A took a situational test and was assessed as "Weak".
They took a general aptitude (IQ) test and got a score of 85.
They describe themselves as someone with high integrity who is focused on the job.

Appendix A

FD2 Candidate A took a situational test and was assessed as "Weak".
They took a general aptitude (IQ) test and got a score of 85.
They describe themselves as someone with high integrity who is focused on the job.

Candidate A took a personality test and scored:
- 3/10 on "openness to experience",
- 5/10 on conscientiousness,
- 2/10 on extroversion and
- 8/10 on agreeableness.

ND1 Candidate A took a situational test and was assessed as "Weak".
They took a general aptitude (IQ) test and got a score of 85.
The candidate scored 5/10 on conscientiousness.

Candidate A loves playing video games and goes to the gym about three times a week. Their favourite sports team is Arsenal.

ND2 Candidate A loves playing video games and goes to the gym about three times a week. Their favourite sports team is Arsenal.

Candidate A took a situational test and was assessed as "Weak".
They took a general aptitude (IQ) test and got a score of 85.
The candidate scored 5/10 on conscientiousness.

Candidate P

DO1 Candidate P took a general aptitude (IQ) test and got a score of 79.
The candidate scored 5/10 on conscientiousness.
They took a situational test and was assessed as "Poor".

DO2 Candidate P took a general aptitude (IQ) test and got a score of 79.
The candidate scored 5/10 on conscientiousness.
They took a situational test and was assessed as "Poor".

FD1 Candidate P took a personality test and scored:
- 8/10 on "openness to experience",
- 5/10 on conscientiousness,
- 7/10 on extroversion and
- 4/10 on agreeableness.

Candidate P took a situational test and was assessed as "Poor".
They took a general aptitude (IQ) test and got a score of 79.
They describe themselves as cheerful and easy going.

FD2 Candidate P took a situational test and was assessed as "Poor".
They took a general aptitude (IQ) test and got a score of 79.
They describe themselves as cheerful and easy going.

Candidate P took a personality test and scored:
- 8/10 on "openness to experience",
- 5/10 on conscientiousness,
- 7/10 on extroversion and
- 4/10 on agreeableness.

Appendix A

ND1	Candidate P took a general aptitude (IQ) test and got a score of 79. The candidate scored 5/10 on conscientiousness. They took a situational test and was assessed as "Poor". Candidate P loves bird watching. Their favourite colour is blue and their favourite food is Asian.
ND2	Candidate P loves bird watching. Their favourite colour is blue and their favourite food is Asian. Candidate P took a general aptitude (IQ) test and got a score of 79. The candidate scored 5/10 on conscientiousness. They took a situational test and was assessed as "Poor".
Candidate O	
DO1	Candidate O scored 4/10 on conscientiousness. The candidate took a situational test and was assessed as "Weak". They took a general aptitude (IQ) test and got a score of 83.
DO2	Candidate O scored 4/10 on conscientiousness. The candidate took a situational test and was assessed as "Weak". They took a general aptitude (IQ) test and got a score of 83.
FD1	Candidate O took a personality test and scored: - 3/10 on "openness to experience", - 4/10 on conscientiousness, - 5/10 on extroversion and - 5/10 on agreeableness. Candidate O took a situational test and was assessed as "Weak". They took a general aptitude (IQ) test and got a score of 83. They describe themselves as a fast learner with high integrity.
FD2	Candidate O took a situational test and was assessed as "Weak". They took a general aptitude (IQ) test and got a score of 83. They describe themselves as a fast learner whit high integrity. Candidate O took a personality test and scored: - 3/10 on "openness to experience", - 4/10 on conscientiousness, - 5/10 on extroversion and - 5/10 on agreeableness.
ND1	Candidate O scored 4/10 on conscientiousness. The candidate took a situational test and was assessed as "Weak". They took a general aptitude (IQ) test and got a score of 83. Candidate O is into martial arts. Their favourite food is Thai and they go to the cinema at least once a month.
ND2	Candidate O is into martial arts. Their favourite food is Thai and they go to the cinema at least once a month. Candidate O scored 4/10 on conscientiousness. The candidate took a situational test and was assessed as "Weak". They took a general aptitude (IQ) test and got a score of 83.

Appendix A

Candidate T	
DO1	Candidate T took a situational test and was assessed as "Poor". They took a general aptitude (IQ) test and got a score of 78. The candidate scored 3/10 on conscientiousness.
DO2	Candidate T took a situational test and was assessed as "Poor". They took a general aptitude (IQ) test and got a score of 78. The candidate scored 3/10 on conscientiousness.
FD1	Candidate T took a personality test and scored: - 5/10 on "openness to experience", - 3/10 on conscientiousness, - 5/10 on extroversion and - 5/10 on agreeableness. Candidate T took a situational test and was assessed as "Poor". They took a general aptitude (IQ) test and got a score of 78. They describe themselves as someone without prestige and a team player.
FD2	Candidate T took a situational test and was assessed as "Poor". They took a general aptitude (IQ) test and got a score of 78. They describe themselves as someone without prestige and a team player. Candidate T took a personality test and scored: - 5/10 on "openness to experience", - 3/10 on conscientiousness, - 5/10 on extroversion and - 5/10 on agreeableness.
ND1	Candidate T took a situational test and was assessed as "Poor". They took a general aptitude (IQ) test and got a score of 78. The candidate scored 3/10 on conscientiousness. Candidate T loves having dinner with friends. In the weekend they go to the park with their dog. They grew up in a rural town as an only child in the family.
ND2	Candidate T loves having dinner with friends. In the weekend they go to the park with their dog. They grew up in a rural town as an only child in the family. Candidate T took a situational test and was assessed as "Poor". They took a general aptitude (IQ) test and got a score of 78. The candidate scored 3/10 on conscientiousness.

Appendix B: Mean rating per candidate and condition, with mean standard deviations.

	DO	FD	ND
M	95.77 (5.54)	75.7 (18.45)	92.68 (8.47)
N	91.79 (8.2)	67.06 (18.86)	89.89 (8.87)
B	94.48 (6.89)	77.1 (17.14)	90.57 (9.38)
C	83.12 (10.86)	76.64 (15.26)	82.96 (11.33)
L	81.58 (9.97)	67.54 (19.19)	80.02 (10.14)
J	77.21 (10.06)	73.56 (15.57)	75.17 (10.7)
H	67.25 (11.28)	71.02 (14.32)	69.53 (11.66)
G	74.73 (10.47)	67.62 (13.98)	72.13 (11.89)
F	63.54 (15.91)	60.54 (15.25)	65.7 (11.25)
D	56.33 (14.78)	56.1 (17.28)	62.02 (14.61)
S	57.79 (16.66)	60.48 (16.31)	55.15 (16.84)
A	35.73 (18.11)	38.62 (20.04)	39.79 (17.68)
P	31.46 (18.99)	42.1 (21.23)	36.15 (19.71)
O	28.73 (15.56)	34.72 (20.77)	31.83 (15.27)
T	19.81 (13.56)	31.3 (18.8)	27.09 (18.2)