Organizational Behavior and Human Decision Processes 137 (2016) 27–44

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring decisions

Edgar E. Kausel [a,b,*], Satoris S. Culbertson [c], Hector P. Madrid [a]

[a] School of Management, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile
[b] Depto. de Administración, Universidad de Chile, Chile
[c] Pamplin School of Business, University of Portland, Portland, OR 97203, United States

ARTICLE INFO

Article history:
Received 28 May 2015
Revised 26 July 2016
Accepted 28 July 2016
Available online 8 August 2016

Keywords:
Judgment and decision making
Behavioral decision theory
Overconfidence
Hiring decisions
Personnel selection
Human resource management
Conscientiousness
General mental ability
Unstructured interviews
Evidence-based management

ABSTRACT

Overconfidence is an important bias related to the ability to recognize the limits of one's knowledge. The present study examines overconfidence in predictions of job performance for participants presented with information about candidates based solely on standardized tests versus those who also were presented with unstructured interview information. We conducted two studies with individuals responsible for hiring decisions. Results showed that individuals presented with interview information exhibited more overconfidence than individuals presented with test scores only. In a third study, consisting of a betting competition for undergraduate students, larger overconfidence was related to fewer payoffs. These combined results emphasize the importance of studying confidence and decision-related variables in selection decisions. Furthermore, while previous research has shown that the predictive validity of unstructured interviews is low, this study provides compelling evidence that they not only fail to help personnel selection decisions, but can actually hurt them.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

What would I eliminate if I had a magic wand? Overconfidence.
[Daniel Kahneman (Shariatmadari, July 15, 2015)]

Extant research on how managers make decisions in personnel selection falls under two main areas. Some researchers have studied lack of bias (e.g., Dipboye, 1982; Lee, Pitesa, Thau, & Pillutla, 2014). Other researchers have conducted policy capturing studies to assess how managers weigh different predictors or interview dimensions (e.g., Dougherty, Ebert, & Callender, 1986; Lievens, Highhouse, & De Corte, 2005). However, there is surprisingly little published research combining what Hammond (1996) calls *external correspondence* (i.e., accuracy) with *internal coherence* (see also Yates, 1982) in judgment and choice research within employee selection contexts. More specifically, little research has examined how the way managers combine information can affect their predictions of job performance when making hiring decisions. This is unfortunate, as job performance is one of the most important variables in organizational behavior and is critical for organizational success (Bowen & Ostroff, 2004).

Similarly, subjective probability or confidence in one's judgment (Harvey, 1997; Hastie & Dawes, 2009; Klayman, Soll, González-Vallejo, & Barlas, 1999) is considered a key construct in the cognitive and decision sciences (Ratcliff & Starns, 2013). As Pleskac and Busemeyer (2010) argued, "confidence [in one's judgment] has long been a measure of cognitive performance used to chart the inner workings of the mind" (p. 864). Yet, confidence is mostly ignored by personnel selection researchers.[1] Personnel selection processes often result in a choice among candidates, and managers' confidence in their decisions is likely to be linked to the type of job offer as well as subsequent events in the selection process. For example, if a manager is certain that the candidate will be a top performer, the manager will be more likely to make an offer with a high salary and attractive perquisites. If the selection and recruitment processes are intertwined (e.g., managers who assess a set of candidates and decide to either make an offer or recruit more

* Corresponding author at: School of Management, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile.
E-mail addresses: ekausel@uc.cl (E.E. Kausel), culberts@up.edu (S.S. Culbertson), hpmadrid@uc.cl (H.P. Madrid).

[1] An exception is Highhouse (2008) who implicitly emphasized the importance of confidence when he noted that people, when thinking about hiring decisions, fail to view them as subject to error (he termed this *the irreducible unpredictability* in personnel selection).

http://dx.doi.org/10.1016/j.obhdp.2016.07.005
0749-5978/© 2016 Elsevier Inc. All rights reserved.

applicants; Seale & Rapoport, 1997), a confident decision maker may be more likely to hire a candidate and terminate the process.

Related to confidence is overconfidence (Soll, Milkman, & Payne, 2015), which has been defined by Koriat, Lichtenstein, and Fischhoff (1980) as "an unwarranted belief in the correctness of one's answers" (p. 108). Decision makers should not only know about facts and relationships between concepts, but also understand the boundaries of their knowledge (Kahneman, 2011; Mannes & Moore, 2013). Russo and Schoemaker (1992) argued that the key issue in overconfidence is metaknowledge: Appreciating what we know and what we do not know. Metaknowledge—and its cousin, self-knowledge—is a value at the heart of many philosophical and religious perspectives (Gertler, 2015). Beyond a value *per se*, a lack of metaknowledge or extreme overconfidence is related to excessive risk; thus, it has consequences on a number of decision-related outcomes (see Goodie, 2003; Malmendier & Tate, 2015; Picone, Dagnino, & Minà, 2014).

Our main goal, therefore, is to study overconfidence among hiring managers when they generate predictions regarding applicants' performance. A second, related goal is to examine whether the ways in which managers combine information about unstructured interviews and other predictors can hurt selection decisions. We focus on the overconfidence of managers presented with information about standardized tests vis-à-vis ratings on unstructured interviews. For the standardized tests, we include measures of conscientiousness (a trait from the Five Factor Model of personality) and general mental ability (GMA). We chose these three predictors (GMA, conscientiousness, and unstructured interviews) because of their frequent use in personnel selection (Farr & Tippins, 2010).

In a set of three studies, we make three contributions to the literature. We first build on previous research involving undergraduate students that suggests unstructured interviews adversely impact predictions of others' performance (Dana, Dawes, & Peterson, 2013). Based on these prior findings, and on research suggesting that GMA and conscientiousness tests are important predictors of job performance (Schmitt, 2014), we expected that experienced managers presented with information of unstructured interviews would have decreased accuracy compared to those presented only with standardized tests. Thus, our first contribution expands previous work to an applied sample. A second contribution of our study is the analysis of potential mechanisms of the above effect. In order to do this, we study different decision-related measures that are important in JDM: judgmental consistency, coefficients of cue utilization, and judgment slope or discrimination. A final contribution of our study is that we highlight the importance of confidence and overconfidence in personnel decisions. A heightened overconfidence, we argue, can have deleterious consequences for decision makers, such as lower financial returns. As such, we argue that personnel selection scholars and practitioners should pay closer attention to confidence and overconfidence, as researchers in other areas have done successfully (e.g., weather forecasting, Tetlock & Gardner, 2015).

The theoretical background for the present study is organized as follows. First, we briefly explain the lens model (Brunswik, 1956) and discuss the expected effect of information presented on accuracy. Then, we explain different decision-related measures that could serve as mechanisms explaining the information-accuracy relationship. Next, we explicate the expected effect of information presented on confidence and overconfidence. Finally, we explain the importance of slope in judgment analysis and state our research question related to this construct.

1.1. Effect of predictor types on accuracy of performance estimates

A useful framework to understand why presenting information on unstructured interviews to practitioners may limit their accuracy in the presence of other (more valid) information is Brunswik's lens model (Brunswik, 1956; Kaufmann, Reips, & Wittmann, 2013; Kuncel, Klieger, Connelly, & Ones, 2013). There are three main components in the lens model (see Fig. 1): the decision maker's judgment, the cues, and the criterion. There are also two main relationships: the relationship between the cues and the criterion (akin to the idea of criterion-related validity) and the relationship between the decision maker's judgment and the cues (i.e., the coefficients of utilization or how the decision maker weighs the different cues). Thus, judgmental accuracy will be high if there is a match between the criterion-related validity and cue weighing. If the *external world* shows that the relationship between GMA and job performance is high and the relationship between unstructured interviews is almost zero, then the decision maker (in his or her *internal world*) should place a high weight on GMA and little to no weight on the interview.

With regard to the left side of the lens, a considerable amount of research has been conducted examining the criterion-related validity of GMA tests, conscientiousness tests, and employment interviews. Perhaps the most consistent finding is that GMA is one of the best predictors of job performance (Ones, Dilchert, Viswesvaran, & Salgado, 2010; Schmitt, 2014). In addition to GMA, two predictors have received considerable attention in the workplace, both due to their predictive capabilities and potential to lessen the adverse impact associated with GMA: Conscientiousness tests and employment interviews. Among the Big Five personality variables, conscientiousness has consistently been shown to be the best predictor of job performance across all occupational groups and job-related criteria (Barrick & Mount, 2012; Barros, Kausel, Cuadra, & Díaz, 2014; Hough & Dilchert, 2010; Hough & Oswald, 2008). In their comprehensive meta-analysis, Barrick, Mount, and Judge (2001) found that the corrected relationship between conscientiousness and job performance was 0.31 (uncorrected, $r = 0.15$).

The interview has also gained substantial attention as a selection tool. One of the key findings that has emerged from numerous meta-analyses (e.g., Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994) is that increased structure (i.e., question and response evaluation standardization) has been associated with higher criterion-related validity of the employment interview. Recent estimates show that whereas the corrected criterion-related validity for unstructured interviews is only 0.20 (uncorrected $r = 0.07$), it is as high as 0.69 (uncorrected $r = 0.36$) for highly structured interviews (Huffcutt, Weyhrauch, & Culbertson, 2014). In terms of incremental contributions, Cortina, Goldstein, Payne, Davison, and Gilliland (2000) found that "unstructured interviews contribute very little, even under ideal circumstances, and interviews high in structure contribute as much, if not more, to prediction as do cognitive ability scores" (p. 340). This conclusion is in line with Schmidt and Hunter's (1998) findings that the combined validity of job performance and unstructured interviews (corrected) is only 0.55 (a slight improvement over GMA's validity coefficient of 0.51). These results suggest that using unstructured interviews to make predictions of job performance when other valid predictors are available is unwise.[2]

Thus, with regard to Brunswik's (1956) model, the left side of the lens is fairly straightforward. If the predictors from which to choose are standardized tests (GMA and conscientiousness tests) and unstructured interviews, in order to match the external world, the decision maker should place some combination of consistent

[2] Some researchers argue that the coefficients involving unstructured interviews in meta-analyses are likely to be overestimated. This is because unstructured interviews are not typically scored (and therefore unavailable for inclusion in a meta-analysis). Those included in a meta-analysis are likely on the high end of rigor (Highhouse, personal communication, May 18, 2012).

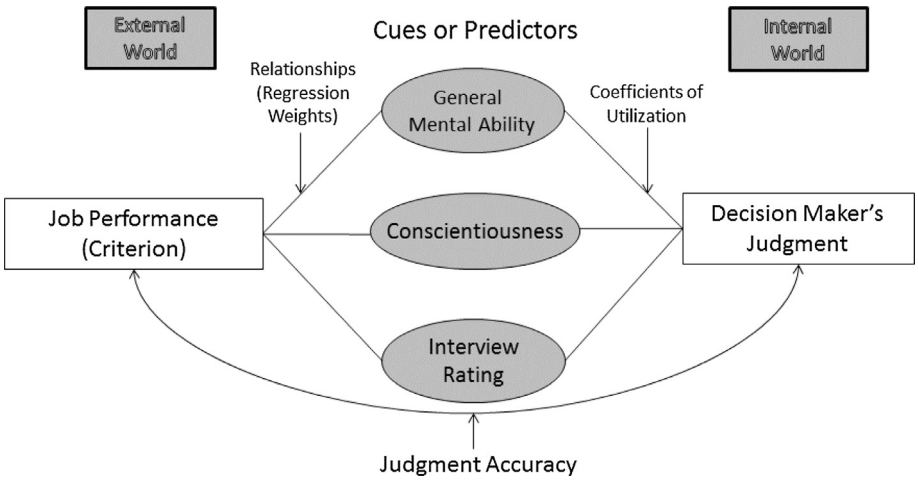

Fig. 1. Lens model applied to employee decisions.

weights on GMA and conscientiousness—which are widely considered as valid predictors of job performance—and little to no weight on unstructured interviews.

In order to further explain why the presence of unstructured interview ratings would decrease accuracy in prediction, we must turn our attention to the right side of the lens, which concerns how people place emphasis on these cues. Terpstra (1996) asked a sample of human resource managers about the validity (perceived effectiveness) of GMA tests, personality tests, and unstructured interviews in making personnel decisions. Perhaps surprisingly, Terpstra found the *opposite* pattern to that suggested by research above. That is, people perceived unstructured interviews as the most effective method for selecting people, personality tests as somewhat less effective, and GMA tests as having the worse perceived effectiveness. This is consistent with Lievens et al.'s study (2005), who examined how managers weighted different information to estimate the hirability of job candidates. The information presented to managers was collected through two different methods: paper-and-pencil tests vs. unstructured interviews. Lievens et al. found that participants placed greater weight on the information collected through unstructured interviews than they did for the information collected through paper-and-pencil tests. Thus, in Brunswik's (1956) terms, most practitioners' internal worlds do not match the external world—the importance they place on different predictors is inconsistent with research findings (for more about this involving other findings in human resource management, see Rynes, Brown, & Colbert, 2002; Terpstra & Rozell, 1997).

Both the lens analogy and these findings suggest two things. First, unsurprisingly, when people are presented with valid cues, it is likely that they will make good decisions. Second, and less obvious, when people are presented with valid *and* invalid cues (e.g., unstructured interview information in addition to GMA and conscientiousness scores), they will make worse decisions than when presented with only the information from the valid predictors because they distribute the cues' weights erroneously. Paradoxically, this implies that having more information could at times hurt selection decisions. This is consistent with Nisbett, Zukier, and Lemley's (1981) findings that non-diagnostic information weakens the validity of diagnostic information in people's predictions, a bias they labeled the *dilution effect*. This is also in line with Dana et al.'s (2013, Study 1) results (see also Bartlett & Green, 1966; Hall, Ariss, & Todorov, 2007; Sarbin, 1943). Dana et al. (2013) studied how undergraduate students changed their predictions of other students' academic performance on the basis of unstructured interviews. Dana et al. provided their participants with information of the target students' grade point average (GPA) from a previous semester. Some of the participants also conducted unstructured interviews; others did not. Dana et al. found that students who conducted these interviews were worse at predicting GPA than those who did not have this information. In essence, the unstructured interview information "diluted" the predictive validity of GPA scores: Had these students simply predicted that future GPA would be equal to past GPA (a predictor that they were given), they would have been more accurate.

With these points in mind, we present the following hypothesis:

Hypothesis 1. Decision makers presented with standardized test scores alone will have greater accuracy in their predictions of subsequent applicant performance than will decision makers presented with standardized test scores and unstructured interview ratings.

We also sought to disentangle the mechanism that explains the effect of presenting unstructured interview information. In order to do this, we take a closer look at the right side of the lens model. A first potential mechanism is judgmental consistency, which is similar to the notion of reliability. Judgmental consistency refers to whether decision makers use information in a similar way across different predictions (Stewart, 2001). Consistent decision makers, when presented with identical information in different occasions, tend to make similar decisions. Just as with test scores, reliability is a necessary, but not sufficient, condition for the validity of judgments. Introducing interview information could add difficulty, and this often decreases consistency. As such, one explanation of Dana et al.'s (2013) results—and more broadly, of the dilution effect—is that conducting or watching interviews (or engaging in any procedure involving vividness) can generate an emotional burden or information overload. As a consequence, participants are less consistent in their estimates, which in turn, reduces their accuracy.

Despite this possibility, we argue that what drives this effect is not inconsistency. Rather, the mechanism is the importance given to less valid information—or, stated differently, the reduced weight that the decision makers put on the valid cues. Individuals who are presented with interview ratings are unable to ignore this invalid information. Unstructured interviews seem to have special allure for hiring managers (Van der Zee, Bakker, & Bakker, 2002) because they seem to tap *psychological factors* (Highhouse, 2008) that allegedly cannot me measured through a different method. As a result, managers make poor decisions. Indeed, those presented with

unstructured interview ratings but who choose to ignore such invalid information could do well in their predictions. However, most people tend to place an unwarranted weight on unstructured interview information, and thus the presence of this cue explains the impaired decisions. Thus, our hypothesis is as follows.

Hypothesis 2. The relationship between information presented (standardized test scores alone vs. standardized test scores and interview ratings) and accuracy will be mediated by use of valid cues. When unstructured interviews are not presented, decision makers will use to a greater degree valid cues (GMA and conscientiousness tests) to make their estimates, which in turn will increase their accuracy.[3]

1.2. Effect of predictor types on confidence of performance estimates

In addition to decreasing accuracy, we also expect that the presence of information about unstructured interviews will increase decision makers' confidence. Not only do practitioners believe that unstructured interviews are useful, but they also believe that they are *more useful* than standardized tests. As Huffcutt and Culbertson (2010) noted in their discussion of why employment interviews remain so pervasive despite other predictors being more reliable and valid, "It is almost as if a part of the human make-up does not trust objective information completely even if it is accurate, the result of which is an underlying desire for personal verification" (p. 185). Practitioners seem to believe that the combination of a standardized test and an intuitive method is very useful, because the first seem to cover "technical competence," while the latter seem to expose the "person as a whole" (Highhouse, 2002; Kuncel & Highhouse, 2011). Furthermore, they believe that interviews and intuition tend to capture all the variance in success that is not attributable to technical competence, including error or simply unpredictable variance (Highhouse, 2008). As a result, they are likely to become more confident when they are presented with standardized test scores *and* interview ratings, as opposed to when they are presented with standardized tests alone.

Based on this rationale, we present the following hypothesis:

Hypothesis 3. Decision makers presented with standardized test scores alone will have less confidence in their predictions of subsequent applicant performance than will decision makers presented with standardized test scores and unstructured interview ratings.

Overconfidence is operationally defined as the subtraction of objective accuracy from subjective confidence[4] (Klayman et al., 1999). As such, based on Hypotheses 1 and 3, a corollary hypothesis is the following:

Hypothesis 4. Decision makers presented with standardized test scores alone will be less biased towards overconfidence than decision makers presented with standardized test scores and unstructured interview ratings.

1.3. Discriminating right and wrong predictions into different confidence levels

Overconfidence is important, but not the only decision-related measure. Another significant factor in judgment and choice analysis is whether people can discriminate correct and incorrect predictions into different confidence levels. More specifically, it is important whether they can assign higher subjective probabilities when their predictions are correct than when they are incorrect (Ronis & Yates, 1987; Whitecotton, 1996). This measure is mostly independent from overconfidence, although it is obtained from the same variables (confidence and accuracy). We wanted to assess whether adding unstructured interview information, while increasing overconfidence, could be beneficial when using a different decision-related outcome.

As such, we used a construct known in the judgment analysis literature as the *slope* (Yates, 1990; also called *separation*, Önkal, Yates, Simga-Mugan, & Öztin, 2003; or *discrimination*, Bonham & González-Vallejo, 2009). This conceptually indicates "the respondent's metacognitive assessment about whether they are right or just guessing" on individual choices (Sieck & Arkes, 2005, p. 34). Operationally, the slope is obtained by subtracting the average confidence assigned to incorrect choices from the average confidence assigned to correct choices. The slope also allows answering the following question: When decision makers are more confident, do they also tend to be more accurate? For this to occur, the available information needs to be valid, but decision makers also need to recognize that they have valid information (Brenner, Griffin, & Koehler, 2005).

Individuals who are high on the slope index discriminate better between their choices in terms of confidence in their decisions (Siegel-Jacobs & Yates, 1996; Vredeveldt & Sauer, 2015; Yates, 2010). For example, decision makers may often choose the right answer but also be unsure about these answers (i.e., on a confidence scale ranging from 50% to 100%, they always choose 50%). In this case, their overconfidence bias may be close to zero, but they would have nil slope (see Dahl, Allwood, Scimone, & Rennemark, 2015). Ideally, decision makers should assign a confidence level of 100% to correct choices, and a confidence level of 50% to incorrect choices, which would result in perfect slope (i.e., in binary decisions, an index of 0.5). This is relevant in personnel selection decisions, beyond overconfidence. Clients of an executive search firm may be unhappy to hear that their advisors always express the same level of confidence across decisions—or that their confidence is simply random (as a slope equal to zero would indicate).

There could be arguments supporting either side on whether the effect of adding unstructured interview ratings to a set of standardized test scores would have a positive or negative effect on the slope index. On the one hand, adding unstructured interview ratings could increase random and systematic error in managers' estimates. Both types of error can negatively affect the slope (Yates, 1990). On the other hand, unstructured interviews tend to correlate with GMA (Cortina et al., 2000). Because GMA is a strong predictor of job performance, it is likely that information from GMA and unstructured interviews would be consistent, which would also increase confidence in correct choices. Thus, even if unstructured interviews do not increase predictive ability (i.e., accuracy), they could help discriminate between high and low accuracy with appropriate levels of confidence. Given the conflicting theorizing regarding this effect, we asked the following research question:

> *Research Question 1: Will decision makers presented with standardized test scores alone have improved* slope *(i.e., discriminate better between correct and incorrect choices) in their predictions of subsequent applicant performance than decision makers presented with standardized test scores and unstructured interview ratings?*

1.4. The present studies

We conducted three studies, in which we manipulated the information presented to participants. In Studies 1 and 2, individuals

[3] Although not included as a formal hypothesis, we do test inconsistency as an alternative mechanism.

[4] Moore and Healy (2008) made a distinction among different types of overconfidence. They suggest that research using a method like the one we used in our studies (*forced choice, half range* tasks, see below) taps both 'overestimation' and 'overprecision' in their taxonomy.

responsible for hiring decisions were presented with scores from actual applicant data and were asked to make decisions regarding which applicants (from paired comparisons) would be the better performers. In Study 1, half of the participants were presented with standardized test scores (GMA and conscientiousness); the other half were presented with standardized test scores (GMA and conscientiousness) *and* unstructured interview ratings. In other words, in Study 1, one group of participants had more information than the other, allowing for a strong test of the negative effect of unstructured interviews on the accuracy of decision makers' estimates. In Study 2, half of the participants were presented with standardized test scores (GMA and conscientiousness) as predictors, as in Study 1; the other half were presented with GMA test scores and unstructured interview ratings. In Study 2, then, the amount of information presented in both conditions was equal (what varied across conditions was conscientiousness vs. unstructured interview ratings). This allowed us to test whether the effect on accuracy, confidence, overconfidence and slope is due to the interview information, as opposed to the amount of information presented. Finally, in Study 3, we used undergraduate students and the design was identical to Study 1 (Information presented: GMA and conscientiousness vs. GMA, conscientiousness, and unstructured interview). However, in addition to asking subjective probabilities, we asked participants to bet on the candidates for the chance to win a prize. This allowed us to rule out the possibility that the findings from the first two studies were driven by a difficulty in understanding subjective probabilities (Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000). Further, the addition of the bet allowed us to examine the impact of overconfidence on relevant outcomes for decision makers.

2. Study 1

2.1. Method

2.1.1. Participants and procedure

Participants in Study 1 consisted of 132 individuals responsible for hiring decisions (65 male; 67 female). Participants were recruited in one of two ways. First, organizational decision makers were recruited during a two-day campus-wide career fair at a large university in the Midwestern United States. During the career fair, members of the research team approached individuals who were representing their organizations while they were working their respective booths. Individuals were only approached when they were not actively engaged with any students so as to not disrupt their work. Individuals were asked to participate in a study examining how organizational decision makers make selection predictions and decisions using limited information. Those interested in participating were given a survey packet that included an informed consent form that assured them that their participation was voluntary and that their responses would remain confidential. In addition, they were given a pre-addressed envelope to mail it back in the event that they were not available when a member of the research team returned to collect the survey. A total of 42 participants were recruited using this method.

The second means of recruitment involved targeted snowball sampling in which 35 students in an MBA course served as participant recruiters in exchange for course credit. Student-recruiters were provided with an email invitation that they distributed to working adults they personally knew who were responsible for making hiring decisions. Recipients of the invitation emails were asked to complete the online survey, available through a link within the email. Student recruiters received nominal extra credit for their involvement. In addition, all students were provided with alternative means of earning the extra credit to prevent them from being tempted to provide false data that could jeopardize the integrity of the data. Targeted snowball sampling of this nature allows for diversity in terms of job type, and has been successfully employed in other organizational studies (cf. Martins, Eddleston, & Veiga, 2002; Matthews & Barnes-Farrell, 2010). A total of 90 participants were recruited using this method.

The two recruitment strategies were compared to determine whether differences in the variables of interest (accuracy, confidence, overconfidence, and slope) existed based on recruitment method. No significant differences between the samples emerged (all $ps > 0.05$). Furthermore, we did not find demographic differences (gender, age, and tenure) across these subsamples. As such, all results are reported with the full sample of 132 participants. In the final sample, participants ranged in age from 20 to 69, with an average age of 39 years ($SD = 13$). Participants were employed full-time in a wide variety of industries, including agriculture, construction, engineering, financial services, insurance, manufacturing, and transportation with a wide variety of occupational titles including manager (e.g., general manager, branch manager, district manager), recruiter, human resources generalist, and executives (e.g., president, vice president, chief operating officer). The length of time that participants reported being employed in their current jobs ranged from three months to 38 years, with an average tenure of 7 years. In response to the question, "How experienced are you with personnel selection decisions?" the average response on a six-point scale (1 = no experience to 6 = extremely experienced) was 4.6 ($SD = 1.0$). Twenty-one percent reported being "extremely experienced." None of the participants reported having "no experience."

2.1.2. Applicants' data (stimuli)

We first randomly sampled candidates from a pool of 236 actual applicants at an airline company. The firm was validating their selection procedures for the job of a Ticket Agent; they were opening new offices so they had several positions opening over the span of eighteen months. Applicants had taken GMA and conscientiousness tests. They also had been interviewed, via an unstructured interview conducted by a line manager, for the job. The scores from the standardized tests were transformed into percentile ranks for the 236 applicants. The only standard to hire an applicant was that they did not receive the worse possible rating in the unstructured interview (rated on a scale from 1 to 5); these people were eliminated from our database. Approximately three months after being hired, these same individuals were assessed by their supervisors on their overall performance. Supervisors used 8 items to assess a variety of dimensions measuring overall performance, including dimensions such as 'discipline,' 'oral communication,' and 'service orientation.' Coefficient alpha for this job performance measure was 0.73.

Based on these data, for the current study we randomly selected the predictors and performance information linked to 40 individuals, which were in turn randomly paired (resulting in 20 pairs). The only requirement for pairing applicants was that they did not have exactly the same performance rating.

2.1.3. Materials

The materials for this study consisted of survey packets that included demographic information, importance ratings for the different selection methods, and pre-hiring information for ten pairs of applicants. Participants were first given information about the airline company and the selection process described above. They were informed that the firm was validating their selection procedures, that more than 200 applicants were assessed using two different selection tools prior to being hired, and that three months after being hired all of them were assessed by their supervisors in terms of their general performance.

Participants were presented with information on ten pairs of applicants' scores and ratings on the predictors, which were presented in a table. The critical manipulation was the predictors presented to participants. Seventy participants were told two tests were given prior to being hired: a standardized GMA test and a standardized personality test (conscientiousness factor). Sixty-two were told that these two tests were given, in addition to ratings from an unstructured interview conducted by a line manager at the airline company. It is important to note that the applicants (and thus their performance) were the same across conditions, but what varied was the information presented to participants. Participants were provided with the ten pairs of applicants' percentile ranks for the GMA test and the conscientiousness test. To ensure understanding, all participants were given the following information pertaining to percentile ranks, "Percentile is the percentage of individuals who score less than the candidate on that dimension. For example, a percentile score of 50 on the cognitive ability test means that the candidate performed better than 50% of the other individuals, and 49% performed better than s/he did." In addition, individuals who received information on the unstructured interviews were informed that the interview ratings ranged from 2 to 5 because, per company policy, those scoring 1 were not hired. They were further informed that 31% of hired candidates received an interview rating of 2; 30% received a rating of 3; 26% received a 4; and 13% received a rating of 5.

We used a type of response known as *forced choice, half range* (Griffin & Brenner, 2004; Hoch & Loewenstein, 1989). Following each pair of applicants' scores on the two sets of predictors, participants were asked to select which applicant (Candidate A or Candidate B) they thought would perform better for the job.[5] In addition, they were asked to indicate their level of confidence in their prediction, from 50% (absolutely uncertain about the prediction) to 100% (absolutely certain the prediction is correct). Participants were further told that a 50% probability rating indicates that they believe it is just as likely that their prediction is correct as incorrect. This half-range measure is appropriate because people who feel less confident than 50% in the chosen candidate should choose the alternative candidate.

In addition to varying the sets of predictors, the packets also varied in terms of the specific applicants who were being compared. Half of participants made comparisons on one sample of ten pairs of applicants while the other half compared ten different pairs of applicants (20 pairs of applicants in total). We did so because asking participants to make too many choices could lead to a fatigue effect or experimental mortality (attrition); at the same time, we wanted to make sure that our results were not due to a small sample of stimuli.

Finally, to account for order and contrast effects, the order of presentation for each set of applicants varied such that half of the participants rated the ten pairs of applicants in one order (1–10) and the other half rated the same pairs of applicants in reverse order (10–1). Thus, there were eight versions of the packets ($2 \times 2 \times 2$ between subjects): two different sets of prediction samples (Sample 1 and Sample 2), two sets of information presented (Set 1: GMA and Conscientiousness, or, for simplicity, 'tests only' condition; Set 2: GMA, Conscientiousness, and Interview Ratings, or, for simplicity, 'interview' condition), and two orders of presentation (Order 1 and Order 2, the reverse of Order 1).

2.1.4. Dependent variables and mediator

Our main dependent variables were accuracy, confidence, overconfidence, and slope. Accuracy at the trial level was a binary variable (*incorrect answer* = 0; *correct answer* = 1), indicating whether the participant chose the candidate that eventually had the better performance of the two—that is, the candidate who eventually received the highest performance rating from their supervisor.[6] We aggregated accuracy across trials and used proportion correct as its operationalization (Fischhoff & MacGregor, 1982). Confidence was a continuous variable that ranged from 0.50 to 1.00, consistent with the half-range measure described above. We also aggregated confidence across the 10 trials and used average confidence for each participant.

We followed Oskamp (1965), who argued that decision makers are calibrated when their confidence matches their accuracy (see also Ronis & Yates, 1987). Analogous measures (but inversed) are over and underconfidence, which occur when there is a difference between confidence and accuracy: overconfidence exists when the difference is positive and underconfidence when the difference is negative (Lichtenstein & Fischhoff, 1977). Thus, overconfidence was defined as average confidence minus proportion correct (Juslin & Montgomery, 2007), theoretically ranging from −1.0 to 1.0. This index has been profusely used within the calibration and overconfidence literature (Mannes & Moore, 2013) to test a match between confidence and accuracy on normative grounds.

The mediator *use of valid cues* was defined as the weighting participants gave to the standardized test scores. We used *R*, an indicator of the partial contribution of predictors when the outcome is dichotomous, as the index of cue weight (Cooksey, 1996; Garson, 2014), akin to relative importance analysis. In each condition, and for each participant, we computed *R* by conducting a logistic regression for each individual, using the cue values presented in each of the 10 choices as the independent variable, and the choices made by the participant as the dependent variable. We operationalized *use of valid cues* as the sum of the absolute value of the raw *R* derived from GMA and conscientiousness scores.[7] Finally, the slope index was computed using Yates, Lee, Shinotsuka, Patalano, and Sieck's (1998) formula. This is defined as the mean subjective probability given that the categorical choice is correct minus the mean subjective probability given that the categorical judgment is incorrect (Dahl et al., 2015; Sieck & Arkes, 2005).

2.2. Results

2.2.1. Applicants' data

Means, standard deviations, reliabilities, and correlations among applicants' variables are included in Table 1. These correlations are generally consistent with uncorrected correlations reported in meta-analyses, described above. This suggests that these data are consistent with what our participants could encounter in different situations (i.e., whether results based on analyzing these data were similar to findings in the meta-analytic literature). We also ran two regression models, both using job performance as the outcome: the first using GMA and conscientiousness scores as predictors; the second using GMA, conscientiousness, and interview ratings as predictors. Table 2 shows the results of these regression analyses. The regression model showed that

[5] A different method could have been asking participants to predict applicants' individual performance and compare these with the applicants' actual performance. However, this would have introduced an extra difficulty to the study, because we would have had to explain the distribution of performance ratings to participants. In addition, there is much research showing that participants in studies using numerical estimates and confidence intervals tend to show exaggerated overconfidence compared to those in studies using binary choices and confidence levels (i.e., overestimation research; Juslin, Wennerholm, & Olsson, 1999), as we did in our studies.

[6] Recall we only paired the information of candidates that had different performance ratings.

[7] As an anonymous reviewer of this paper noted, this measure could underestimate cue use if the cues are correlated. However, because cues are only weakly correlated, this underestimation, if any, would be minimal.

Table 1
Means, standard deviations, and intercorrelations among applicants' variables.

Variable	*M*	*SD*	1	2	3	4
1. GMA (percentile rank)	0.47	0.29	0.82			
2. Conscientiousness (percentile rank)	0.48	0.29	0.09	0.76		
3. Interview	3.20	1.02	0.12	0.07	–	
4. Job performance	3.16	0.44	0.31	0.24	0.06	0.73

Note. $N = 236$. Correlations > |0.13| are significant at $p < 0.05$; correlations > |0.16| are significant at $p < 0.01$.
Alpha reliabilities presented on diagonal.

Table 2
Hierarchical regression analysis predicting job performance from GMA, conscientiousness, and unstructured interviews among applicants.

	ΔR^2	ΔF	β
Step 1	0.097	25.21***	
GMA			0.31***
Step 2	0.044	11.74***	
GMA			0.29***
Conscientiousness			0.21***
Step 3	0.000	0.02	
GMA			0.29***
Conscientiousness			0.21***
Unstructured interview			0.01

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

conscientiousness significantly predicted job performance ($\beta = 0.21$, $p < 0.01$), over and above GMA ($\beta = 0.29$, $p < 0.01$; see Steps 1 and 2, Table 2); both accounting for 14.1% of the criterion variance. However, Step 3 shows that interview ratings did not incrementally predict ($\beta = 0.01$, $p = 0.89$) variance in job performance over GMA and conscientiousness (see Step 3, Table 2).

2.2.2. Regression-based predictions

We also made mechanical predictions based on linear regression models made on the basis of extant research. This would give us a benchmark as to how well participants perform in their estimates (see Edwards & Berry, 2010; Edwards & Christian, 2014). We used regressions based on meta-analytic information on the relationships among job performance, GMA, conscientiousness, and unstructured interviews (Cortina et al., 2000; Roth et al., 2011). We derived the beta weights for three regression equations using: (a) GMA and conscientiousness scores as predictors, (b) GMA, conscientiousness, and interview ratings as predictors, and (c) GMA and interview ratings as predictors (see Study 2). We did this utilizing the *set.cor* function from the psych package in the *R* statistical software (Revelle, 2014). After deriving the beta weights for both regression models, we obtained predicted scores on performance (the criterion) for each of the 236 individuals. Because we had 40 individuals (20 pairs), we were able to compute an accuracy index for the two regression models. We tested whether the regression model's predicted performance was consistent with actual performance; that is, for each pair, whether the model's choice was the same applicant who eventually received a better performance rating. The accuracy index (proportion correct), therefore, was based on the number of "correct choices" by the regression model for each pair, divided by the number of pairs (in this case, 20).

The accuracy rating for the first regression model (GMA and conscientiousness as predictors) was 0.75; the accuracy of the second model using GMA, conscientiousness, and unstructured interviews was 0.80 (in terms of validity, this is equivalent to $R^2 = 0.36$). Thus, participants had predictor information that could yield better-than-chance accuracy rates.

We also conducted relative importance analyses (Johnson, 2000) for these predictors, which established the optimal weighting of these cues. Relative weights are computed by creating a set of variables that are maximally related to the criterion variable but that are orthogonal to each other, avoiding the problems caused by correlated predictors (Tonidandel & LeBreton, 2011). As Fig. 2 shows, when GMA and conscientiousness tests were the predictors, results suggested that GMA accounted for 65.0% of the explained variance, while conscientiousness accounted for 35.0%. When adding the interview ratings, GMA explained 64.3%, conscientiousness 34.7%, and interviews 1.0% of the explained variance.

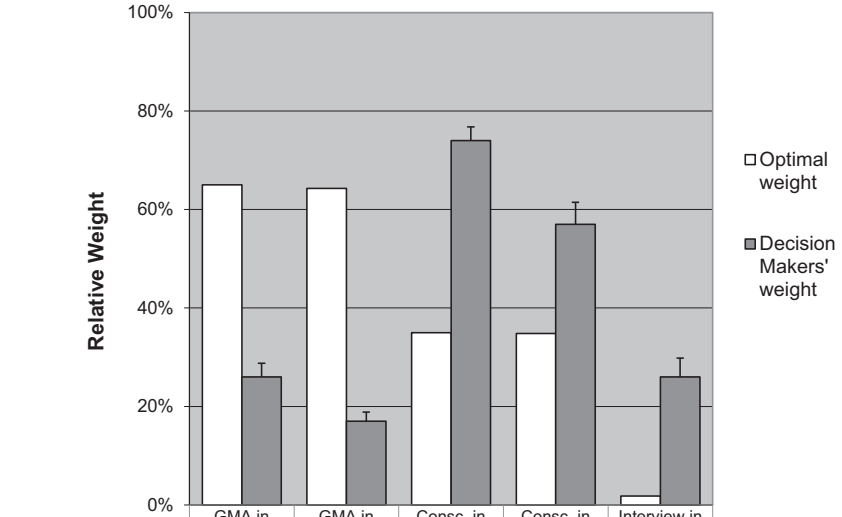

Fig. 2. Optimal and actual relative weight given to cues across conditions, Study 1. *Note.* Optimal weights were obtained from relative weight analyses using the applicants' data set. Error bars represent standard errors of the means. Condition 1: GMA and conscientiousness scores were available; Condition 2: GMA scores, conscientiousness scores, and unstructured interviews ratings were available. Consc. = Conscientiousness.

2.2.3. Testing of hypotheses

Means, standard deviations, and correlations among Study 1 variables are included in Table 3. Results are shown in Fig. 3. Hypothesis 1 proposed that decision makers presented with standardized test (GMA and conscientiousness) scores alone would be more accurate in their predictions than would those presented with standardized test scores and unstructured interview ratings. Results from a regression analysis, using proportion correct as the dependent variable, suggested that the difference was indeed significant, $\beta = 0.27$, $p < 0.01$, $R^2 = 0.07$ (according to Bosco, Aguinis, Singh, Field, and Pierce (2015) this is an effect size at the 67th percentile of effect sizes reported in applied psychology research). Those who based their predictions on standardized tests only had a higher proportion of correct predictions ($M = 0.69$) than those who made predictions based on information that included interview ratings ($M = 0.62$; see Tables 4 and 5). Thus, Hypothesis 1 was supported.

Hypothesis 2 stated that the relationship between information presented and accuracy would be mediated by use of valid predictors, such that decision makers who were not presented with unstructured interviews would be more likely to use valid cues, and in turn would be more accurate. We first computed the weighting that participants gave to each predictor in each of the conditions, employing the R statistic (see above). On the basis of this same analysis, we also computed an average Nagelkerke's R^2 across choices for each candidate as a measure of consistency

Table 5
F-statistics for decision-related variables when the two conditions in each study are compared.

	Study 1	Study 2	Study 3
Accuracy	10.30**	6.10*	7.16**
Confidence	7.54**	5.55*	7.16**
Overconfidence	18.39***	13.79***	15.44***
Slope	0.29	1.18	0.71
Bet amount	–	–	4.11*
Total earnings	–	–	8.66**

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

Table 3
Means, standard deviations, and intercorrelations among Study 1 variables.

Variable	*M*	*SD*	1	2	3	4	5	6
1. Gender	0.49	0.50						
2. Age	39.19	12.58	−0.05					
3. Confidence	0.76	0.09	0.14	−0.14				
4. Accuracy	0.65	0.13	−0.07	0.03	−0.02			
5. Information manipulation	0.52	0.50	0.08	0.02	0.23	−0.27		
6. Hiring experience	4.62	1.04	−0.14	0.27	−0.03	0.10	0.01	
7. Job tenure (months)	88.00	98.38	−0.10	0.57	−0.12	0.03	0.07	0.20

Note. $N = 132$. Gender was coded 0 = female, 1 = male. Information manipulation was coded 0 = interview information not presented, 1 = interview information presented. Correlations > |0.18| are significant at $p < 0.05$; correlations > |0.23| are significant at $p < 0.01$.

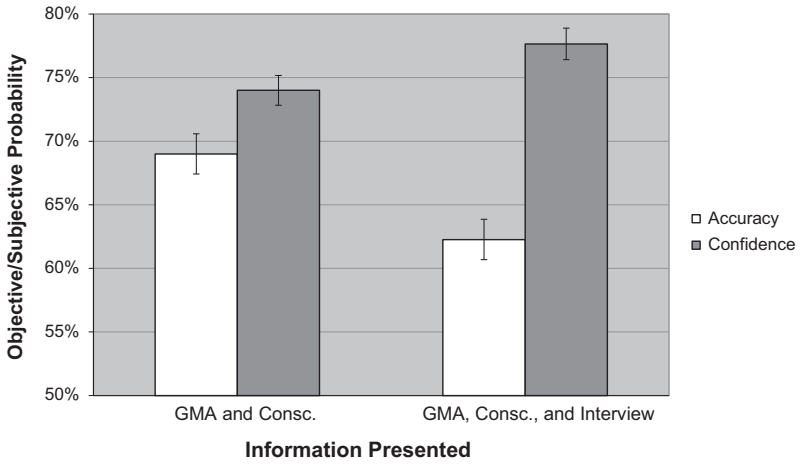

Fig. 3. Accuracy and confidence of decision makers' estimates when presented with different information, Study 1. *Note.* Error bars represent standard errors of the means. Consc. = Conscientiousness.

Table 4
Means (SDs) for decision-related variables across studies.

	Study 1		Study 2		Study 3	
	GMA and consc.	GMA, consc., and interview	GMA and consc.	GMA and interview	GMA and consc.	GMA, consc., and interview
Accuracy	0.69 (0.13)	0.62 (0.12)	0.72 (0.13)	0.63 (0.17)	0.66 (0.13)	0.61 (0.14)
Confidence	0.74 (0.08)	0.78 (0.09)	0.69 (0.12)	0.75 (0.08)	0.77 (0.08)	0.81 (0.09)
Overconfidence	0.06 (0.16)	0.17 (0.13)	−0.02 (0.17)	0.12 (0.16)	0.11 (0.14)	0.21 (0.15)
Slope	0.05 (0.09)	0.05 (0.08)	0.03 (0.11)	0.01 (0.10)	0.05 (0.09)	0.03 (0.10)
Bet amount	–	–	–	–	37.04 (8.75)	39.90 (8.49)
Total earnings	–	–	–	–	143.71 (103.58)	90.75 (115.45)

Table 6
Means (SDs) for weight-related variables across conditions for all studies.

	Standardized tests only condition			Unstructured interview added condition		
	Raw weight	Weight as percentage of R^2	Nagelkerke's R^2	Raw weight	Weight as percentage of R^2	Nagelkerke's R^2
Study 1						
GMA	0.18 (0.12)	0.26 (0.22)		0.14 (0.10)	0.17 (0.13)	
Conscientiousness	0.52 (0.23)	0.74 (0.22)		0.48 (0.18)	0.57 (0.19)	
Interview	–	–		0.22 (0.19)	0.26 (0.14)	
			0.70 (0.18)			0.84 (0.13)
Study 2						
GMA	0.35 (0.21)	0.46 (0.28)		0.36 (0.22)	0.56 (0.30)	
Conscientiousness	0.43 (0.25)	0.55 (0.28)		–	–	
Interview	–	–		0.30 (0.24)	0.44 (0.30)	
			0.78 (0.10)			0.66 (0.18)
Study 3						
GMA	0.11 (0.09)	0.20 (0.16)		0.15 (0.11)	0.22 (0.16)	
Conscientiousness	0.53 (0.24)	0.80 (0.16)		0.41 (0.25)	0.52 (0.26)	
Interview	–	–		0.19 (0.14)	0.26 (0.19)	
			0.64 (0.22)			0.75 (0.14)

(see Cooksey, 1996; Stewart, 2001). The logic here is that if the cues are used consistently (and linearly), then regressing the choices on the cues should lead to a larger R^2. In other words, a higher R^2 implies that decision makers are using the cues consistently.

The 'use of valid cues' variable was higher when participants were presented with GMA and conscientiousness scores, than when they were presented with GMA scores and interview ratings, $t(68) = 8.80$, $p < 0.001$, $d = 2.09$. Results are shown in Fig. 2 and Table 6. To test the mediation, we utilized Hayes' (2012) Model 4 to test multiple mediators. Results of this mediation analysis showed that information presented had a significant indirect effect on accuracy through perceived use of valid predictors (effect = 0.03; 95% CI [0.01, 0.07]), but not through judgment consistency (effect = −0.02, 95% CI [−0.01, 0.06]. This supports Hypothesis 2.

Hypothesis 3 predicted that decision makers presented with standardized test scores alone would have less confidence in their forecasts than those also presented with information about interviews. Regression analysis revealed a significant effect of information presented on confidence, $\beta = -0.23$, $p < 0.01$, $R^2 = 0.06$ (effect size at the 60th percentile; Bosco et al., 2015). Participants in the tests only condition were less confident ($M = 0.74$) in their choices than those in the interview condition ($M = 0.78$). Hence, Hypothesis 3 was supported.

Hypothesis 4 focused on the degree to which participants in the different conditions were overconfident. There was a significant effect of information presented on overconfidence, $\beta = -0.35$, $p < 0.001$, $R^2 = 0.12$ (effect size at the 80th percentile; Bosco et al., 2015). As predicted, decision makers in the tests only condition were less biased towards overconfidence ($M = 0.05$; $SD = 0.16$) than those in the interview condition ($M = 0.16$; see Tables 4 and 5).

Finally, Research Question 1 asked whether judgmental slope (separation or discrimination) would be higher or lower for people who had unstructured interview information versus those who did not. The relationship between information presented and slope was not significant, $\beta = -0.05$, $p = 0.59$. As shown in Tables 4 and 5, managers in the tests only condition were not differentially able to separate or discriminate correct from incorrect predictions (i.e., did not show significantly less slope; $M = 0.05$; $SD = 0.09$) compared to those in the interview condition ($M = 0.05$). This suggests that, although there were differences between confidence and accuracy (as indicated by the differences in overconfidence), these were stable across different levels of confidence.

2.3. Discussion

Results from this study show that, although in both conditions participants were less accurate when compared to a linear model (i.e., regression-based), information derived from standardized tests (GMA and conscientiousness tests) led to greater accuracy than that derived from this same information plus unstructured interview ratings. This was explained by the fact that participants used to a lesser degree valid cues, which predict job performance, when making their choices. In addition, information that included unstructured interview ratings led to greater confidence in individuals' choices. As a result, decision makers who had the extra information of interview ratings to predict job performance were more overconfident than those who only assessed conscientiousness and GMA scores. In terms of slope, there was no difference between people presented with unstructured interview and those presented with standardized tests only. In other words, the presence of unstructured interview information did not significantly improve how decision makers discriminated between accurate and inaccurate choices.

Interestingly, participants underweighted GMA scores in both conditions (see Fig. 2). However, this had different consequences across conditions, as our mediation analyses showed. These results suggest that using conscientiousness to make estimates did not dramatically lower accuracy, whereas using unstructured interview ratings did.

These results are compelling. Individuals who received interview ratings had all the information to be as accurate as those who assessed just standardized tests. However, the inclusion of the interview ratings hurt their selection decisions, by both increasing confidence and decreasing accuracy. Nevertheless, there is a potential confound. Participants who assessed the interview information also had more information. A number of studies have shown that adding information increases confidence levels (Hall et al., 2007; Tsai, Klayman, & Hastie, 2008).

In Study 2 we address this confound. As in Study 1, in one condition individuals received information about standardized tests (GMA and conscientiousness scores). However, in the second condition, we present GMA scores and unstructured interview ratings. Thus, in Study 2, participants received the same amount of information (two predictors) in each condition. By having a controlled study in which the only difference was between the one predictor that was added to GMA test scores (conscientiousness scores vs. unstructured interview ratings), should we find differences, we

can rule out that it was due to information overload or another explanation based on cognitive limitations.

3. Study 2

3.1. Method

3.1.1. Participants and procedure

Participants for this study consisted of 70 individuals responsible for hiring decisions (30 male; 40 female). Similar to one of the recruitment strategies used in Study 1, organizational representatives were approached during a career fair at a large university in the Midwestern United States (in a different state than that of Study 1, and as such no respondents in this study were also in the first study). Similar to Study 1, individuals were approached by a member of the research team and asked to participate in a study examining how selection decisions are made using limited information. Those interested in participating were given a survey packet that included an informed consent form that assured them that their participation was voluntary and that their responses would remain confidential. In addition, they were given a pre-addressed envelope to mail back their survey in the event that they were not available when the researcher returned to collect the survey.

As with Study 1, participants were employed full-time in a wide variety of industries. Participants ranged in age from 22 to 60 years, with an average age of 38 (SD = 12). Job tenure ranged from three months to 30 years, with an average tenure of 6 years and 5 months. In response to the question, "How experienced are you with personnel selection decisions?" the average response on a six-point scale (1 = no experience to 6 = extremely experienced) was 4.6 (SD = 1.2). Twenty-five percent reported being "extremely experienced." None of the participants reported having "no experience."

3.1.2. Applicants' data (stimuli)

The stimuli used for Study 2 were nearly identical to those used in Study 1. The two major differences were that (a) a different set of applicants was sampled, and (b) different cues (predictors) were presented to recruiters. Forty applicants were randomly sampled from the same pool of applicants as in Study 1, which resulted in 20 pairs.

As before, the critical manipulation was the information presented. Thirty-seven participants were shown standardized tests (GMA and conscientiousness scores, the 'tests only' condition), and 33 were shown GMA and unstructured interview scores (the 'interview' condition). Information for 20 applicants (10 pairs) was presented. Participants recorded their responses in a force-choice, half-range format. The two main dependent variables were accuracy (correct choice) and confidence. As in Study 1, we counterbalanced order and used two samples of 20 applicants (10 choices) each.[8]

3.2. Results

We conducted a multiple regression using the applicants' data set, with job performance as the outcome, and GMA, conscientiousness, and interview ratings as the predictors. Interview ratings did not incrementally predict (β = 0.022, p = 0.730) variance in job performance over GMA (β = 0.31, p < 0.01; see Step 2, Table 7). Relative importance analysis suggested that when GMA and unstructured interviews were the predictors, GMA accounted for 98.1% of the explained variance, and unstructured interviews accounted for 1.9%.

Table 7
Hierarchical regression analysis predicting job performance from GMA and unstructured interview among applicants.

	ΔR^2	ΔF	β
Step 1	0.097	25.21***	
GMA			0.31***
Step 2	0.001	0.12	
GMA			0.31***
Unstructured interview			0.02

* p < 0.05.
** p < 0.01.
*** p < 0.001.

Furthermore, the regression-based predictions (based on meta-analytical findings as before), showed that the accuracy (proportion correct) index linked to the model using GMA and conscientiousness as predictors was 0.75; the accuracy of the model using GMA, and unstructured interviews was 0.80.

3.2.1. Testing of hypotheses

Means, standard deviations, and correlations among Study 2 variables are included in Table 8. With regards to Hypothesis 1, the regression analyses showed that information presented had a significant effect on proportion correct, β = 0.29, p < 0.01, R^2 = 0.08. Consistent with this hypothesis, individuals in tests only condition were more accurate (higher proportion of correct predictions; M = 0.72), than those in the interview condition (M = 0.63; see Tables 4 and 5). Hypothesis 1 was supported.

To test the mediation *information presented → use of valid cues → accuracy* (Hypothesis 2), as well as *information presented → consistent use of cues → accuracy* we used a similar procedure as in the previous study. After computing R as an indicator of use of valid cues (see Fig. 4), and Nagelkerke's R^2 as a measure of consistency, we used Hayes' (2012) Model 4 to test for parallel mediators. Hypothesis 2 was supported. Use of valid cues mediated the relationship between information presented and accuracy (indirect effect = 0.20; 95% CI [0.12, 0.30]). In contrast to Study 1, (lack of) consistency also mediated this relationship (indirect effect = −0.02; 95% CI [−0.06, −0.01]). Surprisingly, however, those in the interview condition were more consistent than those in the tests only condition (Table 6 shows details including weights and R^2 across conditions).

As Hypothesis 3 predicted, the regression analyses suggested that there was a significant effect of information presented on average confidence, β = −0.28, p < 0.05, R^2 = 0.08. Participants' confidence in the tests only condition was lower (M = 0.69), than the confidence of those in the interview condition (M = 0.75; see Fig. 5). Furthermore, as stated by Hypothesis 4, there was a significant effect of information presented on overconfidence, β = −0.41, p < 0.001, R^2 = 0.17.

Research Question 1 focused on whether decision makers in different conditions were more confident when they also made accurate predictions (i.e., whether they had different slope). We found no significant differences in the slope of participants in the tests only (M = 0.03) and interview conditions (M = 0.01), β = 0.13, p = 0.28.

3.3. Discussion

Results from Study 2 were generally in line with those from Study 1. First, we found that the regression-based predictions were more accurate than decision makers' estimates. Second, consistent with Study 1, we found that information from unstructured

[8] In two of the eight packets, because of a clerical error, the information involving two choices was missing. Thus, 20 participants made 8 decisions.

Table 8
Means, standard deviations, and intercorrelations among Study 2 variables.

Variable	*M*	*SD*	1	2	3	4	5	6
1. Gender	0.43	0.50						
2. Age	37.84	11.60	0.12					
3. Confidence	0.72	0.10	0.03	−0.08				
4. Accuracy	0.68	0.16	−0.10	−0.03	0.10			
5. Information manipulation	0.47	0.50	−0.10	−0.16	0.28	−0.29		
6. Hiring experience	4.63	1.23	0.07	0.45	0.14	−0.12	−0.18	
7. Job tenure (months)	77.43	77.30	0.15	0.48	−0.11	0.12	−0.05	0.23

Note. $N = 70$. Gender was coded 0 = female, 1 = male. Information manipulation was coded 0 = interview information not presented, 1 = interview information presented. Correlations > |0.24| are significant at $p < 0.05$; correlations > |0.30| are significant at $p < 0.01$.

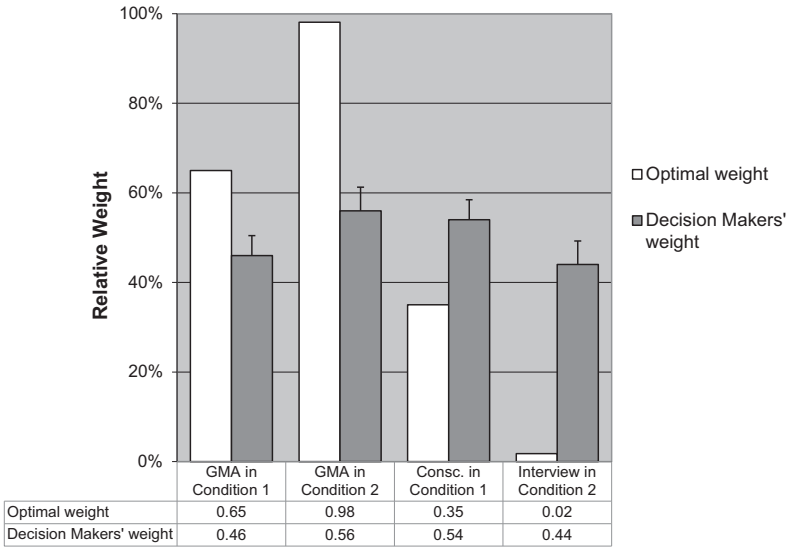

Fig. 4. Optimal and actual relative weight given to cues across conditions, Study 2. *Note.* Optimal weights were obtained from relative weight analyses using the applicants' data set. Error bars represent standard errors of the means. Condition 1: GMA and conscientiousness scores were available; Condition 2: GMA scores and unstructured interviews were available. Consc. = Conscientiousness.

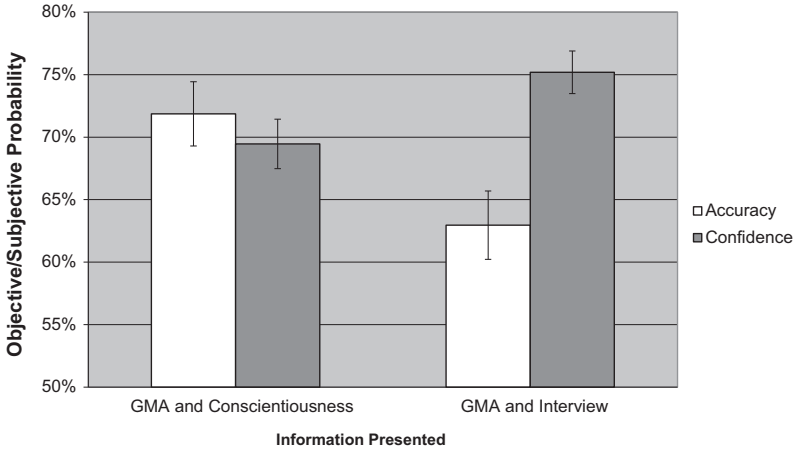

Fig. 5. Accuracy and confidence of decision makers' estimates when presented with different information, Study 2. *Note.* Error bars represent standard errors of the means.

interviews decreased accuracy; and that this was explained in part by the use of valid cues (though consistency had a role as well). Third, including information about unstructured interviews increased confidence. As a result, decision makers who assessed interview ratings and GMA scores to predict job performance were more overconfident than those who assessed conscientiousness and GMA scores. Fourth, we found no significant differences in terms of slope between people who assessed standardized tests only and those who assessed the interview as well. This indicates that the unstructured interview ratings do not help (but also do not hurt) discriminating correct from incorrect decisions.

There was one important difference between the two studies: the specific weights given to the predictors. In Study 2, GMA was given substantially more weight than in Study 1, across conditions. We provide a potential explanation in Section 5. Despite this, in both studies emerged a consistent pattern. Whereas GMA was

Table 9
Means, standard deviations, and intercorrelations among Study 3 variables.

Variable	*M*	*SD*	1	2	3	4	5	6
1. Gender	0.36	0.48						
2. Age	23.88	7.41	0.01					
3. Confidence	0.79	0.08	0.00	0.14				
4. Accuracy	0.63	0.13	0.01	0.06	0.07			
5. Information manipulation	0.51	0.50	0.10	0.08	0.22	−0.22		
6. Bet amount average	38.50	8.71	0.15	0.08	0.69	0.18	0.17	
7. Total earnings	116.70	112.60	0.01	0.08	0.21	0.88	−0.24	0.39

Note. $N = 149$. Gender was coded 0 = female, 1 = male. Information manipulation was coded 0 = interview information not presented, 1 = interview information presented. Correlations > |0.16| are significant at $p < 0.05$; correlations > |0.21| are significant at $p < 0.01$.

given less weight than the optimal, the unstructured interview was given more weight than the optimal.

We conducted a third study with the following aims. First, we wanted to rule out the possibility that findings from Studies 1 and 2 were driven by a difficulty in understanding subjective probabilities (Hoffrage et al., 2000). Second, we sought to replicate the previous findings by giving monetary rewards and punishments to those people who expressed higher confidence in their choices (i.e., we tested whether participants, as some economists like to say, would "put their money where their mouth is"). Related to this, we wanted to examine whether unstructured interview information not only increases overconfidence, but also leads to fewer payoffs.

4. Study 3

4.1. Method

4.1.1. Participants and procedure

Participants were undergraduate students who were recruited from upper-level management and psychology courses at two mid-sized universities in the United States. Participants were provided with extra credit in their classes for participating, and were informed that their participation was voluntary. All data were collected online. We obtained usable data from 149 undergraduate students. The average age of participants was 23 years ($SD = 7.75$) and approximately 65% of them were female.

4.1.2. Materials and stimuli

In Study 3, the design was identical to Study 1 (Information presented: GMA and conscientiousness vs. GMA, conscientiousness, and unstructured interview). The sample of candidates (stimuli) and measures were also the same (with one addition, see below) that were included in Study 1.[9]

There were two differences with Study 1. First, because the task was computer-based, the order of presentation of the candidates was randomized. Second, there was a betting competition. After asking participants for their candidate choice and confidence level, we also asked them to bet on the candidate they chose. On each decision, they could wager any value from $10 to $50. Participants were told that, as part of the competition, for each of the ten decisions they would win the money they bet if they chose the correct candidate (the candidate whose eventual job performance was higher) but would lose the money if they chose the wrong candidate. (Note that these were "points," not actual money; although the participant who ended up with more points would receive a cash prize.) Their final payoff was determined by adding their bets linked to correct choices minus their bets linked to wrong choices. Thus, the potential range of final payoffs was from −$500 (always betting $50 on the wrong candidate) to $500 (always betting $50 on the correct candidate). In order to enhance realism and create an incentive to exert deliberate effort, participants were informed that the top four performers (i.e., those with the highest payoffs) would receive a prize of $50. In the event of a tie, the $50 would be divided between the individuals in the tie.

4.2. Results

Means, standard deviations, and correlations among Study 3 variables are included in Table 9. Note the high correlation between confidence and the 'bet' variable ($r = 0.69$), suggesting that they are measuring essentially the same construct.

4.2.1. Testing of hypotheses

Consistent with Hypothesis 1, individuals in the tests only condition were more accurate ($M = 0.66$) than those in the interview condition ($M = 0.61$), $\beta = 0.22$, $p < 0.01$, $R^2 = 0.05$. Hypothesis 2 was supported: there was an indirect effect of information presented on accuracy through use of valid cues (effect = 0.04; 95% CI [0.00, 0.08]); consistency did not significantly mediate the information presented–accuracy relationship (effect = 0.02; 95% CI [−0.01, 0.06]). (See Table 6 for details on the weights participants gave to each cue.) Hypothesis 3 was also supported, both when measuring confidence in terms of subjective probability or betting behavior. The values of these variables were significantly lower for individuals in the tests only condition ($M_{subjective\ probability} = 0.77$; $M_{betting\ behavior} = 37.03$) than for those in the interview condition ($M_{subjective\ probability} = 0.81$; $M_{betting\ behavior} = 39.9$), $\beta = -0.22$, $p < 0.01$, $R^2 = 0.05$ and $\beta = -0.17$, $p < 0.05$, $R^2 = 0.03$, respectively (see Tables 4 and 5). Those in the tests only condition also showed less overconfidence ($M = 0.11$) than in the interview condition ($M = 0.21$), $\beta = -0.31$, $p < 0.01$, $R^2 = 0.10$. Finally, the answer to Research Question 1 was negative: there were no significant differences in the slope of people in the tests only condition ($M = 0.05$) compared to those in the interview condition ($M = 0.03$), $\beta = 0.07$, *ns*.

4.2.2. Additional analyses

We finally compared the total payoffs earned by participants in the tests only vs. interview conditions. As shown in Tables 4 and 5, those provided only the tests earned significantly more ($M = 143.71$) than those also given interview information ($M = 90.75$), $\beta = 0.24$, $p < 0.01$, $R^2 = 0.10$. We also tested the following mediation: *information presented → overconfidence → total payoffs*. We found a significant effect of information presented on total

[9] We also were initially interested in how outcome feedback affected participants' decisions in different conditions. As such, roughly half of the individuals were given outcome feedback (i.e., they were told which candidate eventually performed better) after each decision. This was distributed across tests only and interview conditions (i.e., a 2 [feedback: yes vs. no] by 2 [information presented: tests only vs. interview] design). However, we did not find main effects of feedback on any of our dependent variables. Nor did we find an interaction between information presented and feedback on these variables. We thus collapsed this variable and do not discuss it further in the paper.

payoffs through overconfidence (effect = 44.20; 95% CI [22.96, 71.13]).

4.3. Discussion

This study provides converging evidence for the results shown in the previous experiments. Presenting unstructured interview information to participants made them (a) less accurate, (b) more confident and more likely to bet more money for their preferred candidates, (c) more overconfident, and (d) not necessarily better able to discriminate among choices (the slope variable). As a result, participants who were provided with unstructured interview information ended up with a lower payoff. Thus, this experiment shows that the addition of unstructured interview information results in worse personnel selection decisions and ultimately lower payoffs.

In addition, Study 3 addressed a common criticism of subjective probability and overconfidence research. This criticism points out that participants typically do not understand information involving probabilities and some of the results may be driven by a lack of understanding (e.g., Hoffrage et al., 2000; Juslin & Montgomery, 2007). We used a more straightforward task, which also provided an incentive to participants to make a more concerted effort to be more accurate.

5. General discussion

Overconfident decision makers are problematic because they have an illusion of understanding (Kahneman, 2011). They lack the metacognitive skill to understand how correct or incorrect their decisions are—and they are willing to take more risks. Furthermore, they fail to acknowledge the importance of uncertainty in personnel selection and that decisions are subject to error (Highhouse, 2008). As a result, overconfident hiring managers may be more willing to extend better job offers than is warranted, or may terminate a selection process before finding the (actual) best candidate.

Given this, our goal was to assess to what degree a sample of hiring managers were overconfident in predicting candidates' performance, as well as the role of the information presented to them had in their estimates. Using a representative design (Hogarth, 2005), results revealed that when individuals assessed information including unstructured interview ratings they were more overconfident than those who assessed information about standardized tests alone. This effect was more than simply judgmental inconsistency, and rather it was due to the inflated weight participants put to the unstructured interview (i.e., failing to weight the valid cues), as the mediation analysis showed. In addition, this overconfidence did not come at a price on another important decision variable: the slope. Those who received unstructured interview information did not discriminate better between right and wrong choices. Furthermore, our results suggested that the overconfidence of those who assessed unstructured interview information resulted in fewer payoffs for these individuals.

An additional finding was that people tended to be less accurate than regression-based predictions. Although this finding is not novel (Swets, Dawes, & Monahan, 2000), the present research constitutes one of the first studies testing this with experienced organizational decision makers who routinely make hiring decisions (beyond the assessment of prospective students' applications to educational programs), and including widely used predictors such as unstructured interviews and tests of conscientiousness and GMA. Furthermore, we derived our regression weights based on available meta-analytical correlation matrices, not on the same applicants' data, which could give an unfair advantage to the regression model. Even when having extra contextual information—such as job type, firm's industry, and type of job performance—the experienced managers were less accurate than the context-free regression model.

Interestingly, a number of authors argue that people, while poor at integrating cues because of lack of consistency (Karelaia & Hogarth, 2008), do well at choosing and weighting predictors (e.g., Kuncel et al., 2013). In fact, a classic recommendation of the JDM literature is to *bootstrap* experts or experienced individuals (derive a model by capturing their preferences) and then using the model to make predictions (e.g., Kaufmann et al., 2013). Our results suggest that experienced individuals do not always weigh cues appropriately, however. The present studies show that these decision makers underestimated the validity of GMA and overestimated the validity of unstructured interviews (see also Lievens et al., 2005; Rynes et al., 2002; Terpstra, 1996).

5.1. Contributions, limitations and future research

Our paper makes a number of contributions. We underscore the importance of assessing subjective probability estimates or confidence in one's judgment in personnel decisions. Study 3 used a betting task and we found that participants who assessed unstructured interview information, because it led them to more overconfidence, ended up with fewer payoffs. This betting task could also be compared to making job offers to candidates. People are more willing to make higher job offers when they are more confident that a candidate will perform well. As such, overconfidence can have important implications to personnel selection outcomes. Related to this, personnel selection scholars and practitioners should take into account confidence and job offers when conducting utility analysis (Guion & Highhouse, 2006). In particular, if managers are more (vs. less confident) that a candidate will be a top performer, a more comprehensive utility analysis could help determining the returns of making a high (vs. a low) job offer. This line of research could have important practical implications.

In addition, while previous research had found that the predictive validity of unstructured interviews is low (Huffcutt et al., 2014), and that their incremental validity over GMA standardized tests is close to zero (Cortina et al., 2000), our study shows that including information about unstructured interviews can actually hurt selection decisions when additional information is also available (see also Dana et al., 2013). In other words, when assessing applicants' credentials, the use of unstructured interviews can lead to a 'more is less effect.' Adding unstructured interview ratings to a set of predictors that already include standardized test scores can increase overconfidence, creating an illusion of knowledge (Hall et al., 2007).

Our results also suggest that, although managers' overconfidence was not particularly large (especially when no interview information was presented to them), the slope of their judgments was quite poor. Across experiments, the highest average of the slope index was 0.05 (Study 1). This is far from the normative slope index, 1.0, which indicates perfect discrimination (Sieck & Arkes, 2005). In other words, decision makers' subjective probabilities tended to be unrelated to their accuracy, within their choices. This is problematic, because hiring managers and especially staffing consultants should have differentiated subjective probabilities depending on their estimated probabilities of the candidates' success.

We should point out some limitations of our paper. A first limitation is that, just as with studies based on surveys or scenarios (Rynes et al., 2002), we focused here on practitioners' *rational* (though inaccurate) belief of the usefulness of unstructured interviews, and studied how this belief affected accuracy and confidence. In other words, our study focused on how decision makers combined cues, but not how they collected information

(Gatewood, Feild, & Barrick, 2010). Managers who conduct interviews—especially unstructured—are often affected by vivid impressions, more than numbers. It is often the case that managers who make the hiring decision also interview the candidates; this is in contrast to our study, in which individuals examined interview ratings made by other managers. Interviews likely convey detailed, individuating, and emotional information about the candidates, factors that are in general heavily weighted (Kahneman & Tversky, 1973). Thus, *experiencing* interviews (not only analyzing others' ratings) can lead practitioners to place a great deal of importance on this method. However, this issue suggests that we are underestimating the effects of adding unstructured interview information on overconfidence; our results could have been even stronger if participants had conducted the interviews themselves.

Another limitation is that we did not use an orthogonal design by fully crossing values of each cue. Thus, the interpretation of the weights that participants gave to the cues is less robust (Karren & Barringer, 2002). This, and the fact that we used a reduced number of decisions to keep the survey short for managers, may explain the difference between cue weights in Study 1 (and 3) versus Study 2. We decided not to use an orthogonal design because we were willing to sacrifice robustness of weights in exchange of representativeness and practicality (Aiman-Smith, Scullen, & Barr, 2002; Highhouse, 2009). Representativeness and practicality were crucial for our studies. Related to this point, although we were careful in the design of our studies to control for alternative explanations, it is important to consider the issue of dominating alternatives (Tenbrunsel & Diekmann, 2002). When individuals are presented with alternatives in which there is a dominant choice, they are more confident in making their decisions than when an alternative does not clearly dominate the other. Given that respondents received GMA, conscientiousness, and/or unstructured interview ratings, it was unclear to what extent their confidence estimates mirrored patterns of attribute dominance. To answer this question, we conducted a series of tests to rule out the effect of mere dominance effects. Our results suggested that dominance issues did have an effect; indeed, these dominance effects may explain the different cues' weights found in Studies 1 and 2. However, our findings regarding the impact that unstructured interviews had on overconfidence remained robust. In other words, decision makers have higher confidence in unstructured interviews beyond an explanation of mere dominance (or cues' consistency). The results of these additional analyses are available by request from the first author.

Finally, we used a *forced choice, half range task* to elicit accuracy and confidence. A different method could have been asking participants to predict applicants' individual performance and compare these with the applicants' actual performance. However, this would have introduced an extra difficulty to the study, because we would have had to explain the distribution of performance ratings to participants. In addition, there is much research showing that participants in studies using numerical estimates and confidence intervals (i.e., overprecision research; Moore & Healy, 2008) tend to show more bias compared to those in studies using forced choice and half range tasks (Juslin et al., 1999), as we did in our studies. For this reason, we could be underestimating overconfidence.

There are at least two avenues of future research. First, future researchers would do well to study ways to reduce overconfidence and enhance slope. For example, an intervention that could be adopted among hiring managers is to ask them to make exact predictions (as in the present study) and then provide feedback. Few practitioners make precise estimates of applicants' performance. For example, recruiters or managers could be asked to what degree they believe that a candidate will be in the top decile of all employees' performance (Slaughter & Kausel, 2013). After a year, they would receive feedback of their estimates, including calibration and over/underconfidence indices. (This would apply, of course, only to accepted candidates.) It is important to note, however, that the mere exposure to outcome feedback does not guarantee learning. This has been shown in both overconfidence research and multiple-cue probability learning studies (Pulford & Colman, 1997; Sieck & Arkes, 2005; Sharp, Cutler, & Penrod, 1988). Indeed, as we note in footnote 9, we did provide outcome feedback to some of the participants but found no effect. As such, important issues beyond outcome feedback, such as meta-task processes, should be taken into consideration when providing feedback in future research (see Feedback Intervention Theory; Kluger & DeNisi, 1996). We also believe that presenting summary feedbacks in frequency format, as opposed to correlation coefficients, will greatly enhance the understanding of the information presented to decision makers (Brooks, Dalal, & Nolan, 2014; Kuncel, 2008).

Second, the negative effects of overconfidence can be not only financial, as we showed in the present research, but social as well. Some level of confidence is linked to increased persuasiveness (Zarnoth & Sniezek, 1997); however, when people have information about the accuracy of decisions, confidence can backfire (Tenney, MacCoun, Spellman, & Hastie, 2007). Furthermore, extreme overconfidence implies expressing daring statements, which, combined with negative results, can have deleterious consequences for the decision maker's reputation. If a consultant from a young executive search firm expresses absolute confidence that a candidate will exhibit integrity at work, but the candidate eventually ends up involved in a scandal, it could be a serious blow to the firm's reputation. After catastrophes, bold leaders are usually blamed for having shown overconfidence (e.g., the mountain guides of the 1996 Everest disaster). Research in other forecasting settings has shown that stakeholders value when there is a match between confidence and accuracy in predictions (Sah, Moore, & MacCoun, 2013; Yates, Price, Lee, & Ramirez, 1996). For example, if a manager is informed by an executive search firm that they are 99% confident that they will choose the best candidate, the manager may understandably be upset to learn that when the consultants express this level of confidence, they are in reality correct only 75% of the time. There is some suggestion that extreme confidence by itself can create skepticism in some contexts (cf. Sah et al., 2013). As Yates et al. (1996) suggested, some people may suspect that "extreme forecasters might know so little about the inherent uncertainty in the situation that they fail to realize that they are being reckless" (p. 54). Future researchers could investigate managers' *reactions* to measures such as overconfidence or discrimination (slope).

5.2. Conclusions

The current research underscores the importance of confidence in one's judgment and overconfidence in hiring decisions, an issue that has been mostly ignored by personnel selection research. We found that organizational decision makers were keenly overconfident when predicting applicants' performance based on information including unstructured interview ratings. We also showed that overconfidence can be linked to reduced payoffs; thus, it may have effects on important outcomes such as the utility of the selection process.

Unstructured interviews may be useful to provide information to applicants, to improve applicants' reactions, and to make sense of the process by integrating different organizational stakeholders (Hausknecht, Day, & Thomas, 2004; Klimoski & Jones, 2008; Stevens, 1998). However, if the goal is to hire the best candidate, our findings suggest that managers should avoid unstructured interviews—they can boost overconfidence and reduce financial returns.

Acknowledgments

Financial support from FONDECYT, under Grant Iniciación #11130277, is gratefully acknowledged. We also acknowledge support from Núcleo Milenio Research Center in Entrepreneurial Strategy Under Uncertainty (NS130028). We thank Alejandro Hirmas, Maximilano Escaffi, and Cristian Vasquez for their help with tables, figures, references, and data analysis.

Appendix A

Pairs of candidates used as stimuli, Studies 1 and 3

Pair 1, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	85	95	3
Candidate B	82	09	4

Pair 2, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	85	07	2
Candidate B	66	11	3

Pair 3, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	51	23	5
Candidate B	41	28	2

Pair 4, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	26	73	4
Candidate B	26	35	2

Pair 5, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	36	28	3
Candidate B	46	86	4

Pair 6, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	33	35	5
Candidate B	12	68	5

Pair 7, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	85	95	4
Candidate B	61	28	3

Pair 8, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	93	86	3
Candidate B	09	51	2

Pair 9, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	66	51	5
Candidate B	04	23	2

Pair 10, Sample 1

	GMA	Conscientiousness	Unstructured interview
Candidate A	57	35	3
Candidate B	41	35	4

Pair 1, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	85	95	3
Candidate B	82	09	4

Pair 2, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	85	07	2
Candidate B	66	11	3

Pair 3, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	51	23	5
Candidate B	41	28	2

Pair 4, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	26	73	4
Candidate B	26	35	2

Pair 5, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	36	28	3
Candidate B	46	86	4

Pair 6, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	33	35	5
Candidate B	12	68	5

Pair 7, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	85	95	4
Candidate B	61	28	3

Pair 8, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	93	86	3
Candidate B	09	51	2

Pair 9, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	66	51	5
Candidate B	04	23	2

Pair 10, Sample 2

	GMA	Conscientiousness	Unstructured interview
Candidate A	57	35	3
Candidate B	41	35	4

References

Aiman-Smith, L., Scullen, S. E., & Barr, S. H. (2002). Conducting studies of decision making in organizational contexts: A tutorial for policy-capturing and other regression-based techniques. *Organizational Research Methods, 5*(4), 388–414.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1–2), 9–30. http://dx.doi.org/10.1111/1468-2389.00160.

Barrick, M. R., & Mount, M. K. (2012). Nature and use of personality. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 225–251). New York, NY: Oxford University Press.

Barros, E., Kausel, E. E., Cuadra, F., & Díaz, D. A. (2014). Using general mental ability and personality traits to predict job performance in three Chilean organizations. *International Journal of Selection and Assessment, 22*(4), 432–438. http://dx.doi.org/10.1111/ijsa.12089.

Bartlett, C. J., & Green, C. G. (1966). Clinical prediction: Does one sometimes know too much? *Journal of Counseling Psychology, 13*(3), 267–270.

Bonham, A. J., & González-Vallejo, C. (2009). Assessment of calibration for reconstructed eye-witness memories. *Acta Psychologica, 131*(1), 34–52. http://dx.doi.org/10.1016/j.actpsy.2009.02.008.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*(2), 431–449. http://dx.doi.org/10.1037/a0038047.

Bowen, D. E., & Ostroff, C. (2004). Understanding HRM-firm performance linkages: The role of the "strength" of the HRM system. *The Academy of Management Review*, 203–221.

Brenner, L., Griffin, D., & Koehler, D. J. (2005). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes, 97*(1), 64–81. http://dx.doi.org/10.1016/j.obhdp.2005.02.002.

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology, 99*(2), 332–340. http://dx.doi.org/10.1037/a0034745.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Diego, CA: Academic Press.

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*(2), 325–351. http://dx.doi.org/10.1111/j.1744-6570.2000.tb00204.x.

Dahl, M., Allwood, C. M., Scimone, B., & Rennemark, M. (2015). Old and very old adults as witnesses: Event memory and metamemory. *Psychology, Crime & Law, 21*(8), 764–775. http://dx.doi.org/10.1080/1068316X.2015.1038266.

Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*(5), 512–520.

Dipboye, R. L. (1982). Self-fulfilling prophecies in the selection-recruitment interview. *Academy of Management Review, 7*(4), 579–586. http://dx.doi.org/10.5465/AMR.1982.4285247.

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71*(1), 9–15. http://dx.doi.org/10.1037/0021-9010.71.1.9.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13*(4), 668–689. http://dx.doi.org/10.1177/1094428110380467.

Edwards, J. R., & Christian, M. S. (2014). Using accumulated knowledge to calibrate theoretical propositions. *Organizational Psychology Review, 4*(3), 279–291. http://dx.doi.org/10.1177/2041386614535131.

Farr, J. L., & Tippins, N. T. (2010). *Handbook of employee selection*. New York: Routledge.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1*(2), 155–172. http://dx.doi.org/10.1002/for.3980010203.

Garson, G. D. (2014). *Logistic regression: Binomial and multinomial*. Asheboro, NC: Statistical Associates Publishers.

Gatewood, R., Feild, H. S., & Barrick, M. (2010). *Human resource selection* (7th ed.). Mason, OH: Cengage Learning.

Gertler, B. (2015). Self-knowledge. In *Stanford encyclopedia of philosophy*.

Goodie, A. S. (2003). The effects of control on betting: Paradoxical betting on items of high confidence with low value. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(4), 598.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177–199). Malden, MA: Blackwell Publishing Ltd..

Guion, R. M., & Highhouse, S. (2006). *Essentials of personnel assessment and selection*. Mahweh, NJ: Erlbaum.

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*(2), 277–290.

Hammond, K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York, NY, US: Oxford University Press.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences, 1*(2), 78–82.

Hastie, R., & Dawes, R. M. (2009). *Rational choice in an uncertain world: The psychology of judgment and decision making* (2nd ed.). Los Angeles, CA: Sage Publications Inc..

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639–683. http://dx.doi.org/10.1111/j.1744-6570.2004.00003.x.

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [white paper]*. The Ohio State University. Retrieved from <http://www.afhayes.com/public/process2012.pdf>.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55*(2), 363–396. http://dx.doi.org/10.1111/j.1744-6570.2002.tb00114.x.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1*(3), 333–342. http://dx.doi.org/10.1111/j.1754-9434.2008.00058.x.

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods, 12*(3), 554–566. http://dx.doi.org/10.1177/1094428107300396.

Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 605–619. http://dx.doi.org/10.1037/0278-7393.15.4.605.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 22*, 2261–2262.

Hogarth, R. M. (2005). The challenge of representative design in psychology and economics. *Journal of Economic Methodology, 12*(2), 253–263.

Hough, L., & Dilchert, S. (2010). Personality: Its measurement and validity for employee selection. In J. L. Farr & L. M. Hough (Eds.), *Handbook of employee selection* (pp. 299–319). New York, NY: Routledge.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial–organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology, 1*(3), 272–290. http://dx.doi.org/10.1111/j.1754-9434.2008.00048.x.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*(2), 184–190. http://dx.doi.org/10.1037/0021-9010.79.2.184.

Huffcutt, A. I., & Culbertson, S. S. (2010). Interviews. In S. Zedeck (Ed.). *APA handbook of industrial and organizational psychology* (Vol. 2, pp. 185–203). Washington, DC: American Psychological Association.

Huffcutt, A. I., Weyhrauch, W. S., & Culbertson, S. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with Indirect Range Restriction Methodology. *International Journal of Selection and Assessment, 22*(3), 297–309.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35*(1), 1–19. http://dx.doi.org/10.1207/S15327906MBR3501_1.

Juslin, P., & Montgomery, H. (2007). *Judgment and decision making: Neo-Brunswikian and process-tracing approaches*. Mahwah, NJ: Psychology Press.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1038–1052. http://dx.doi.org/10.1037/0278-7393.25.4.1038.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. http://dx.doi.org/10.1037/h0034747.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404–426.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of the policy-capturing methodology in organizational research: Guidelines for research and practice. *Organizational Research Methods, 5*(4), 337–361.

Kaufmann, E., Reips, U.-D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PLoS ONE, 8*(12), e83528. http://dx.doi.org/10.1371/journal.pone.0083528.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79*(3), 216–247. http://dx.doi.org/10.1006/obhd.1999.2847.

Klimoski, R., & Jones, R. G. (2008). Intuiting the selection context. *Industrial and Organizational Psychology, 1*(3), 352–354.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107–118. http://dx.doi.org/10.1037/0278-7393.6.2.107.

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology, 1*(3), 343–346.

Kuncel, N. R., & Highhouse, S. (2011). Complex predictions and assessor mystique. *Industrial and Organizational Psychology – Perspectives on Science and Practice, 4*(3), 302–306.

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. http://dx.doi.org/10.1037/a0034156.

Lee, S. Y., Pitesa, M., Thau, S., & Pillutla, M. (2014). Discrimination in selection decisions: Integrating stereotype fit and interdependence theories. *Academy of Management Journal, 58*(4), 789–812. http://dx.doi.org/10.5465/amj.2013.0571.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159–183.

Lievens, F., Highhouse, S., & De Corte, W. (2005). The importance of traits and abilities in supervisors' hirability decisions as a function of method of assessment. *Journal of Occupational and Organizational Psychology, 78*(3), 453–470.

Malmendier, U., & Tate, G. (2015). Behavioral CEOs: The role of managerial overconfidence. *Journal of Economic Perspectives, 29*(4), 37–60.

Mannes, A. E., & Moore, D. A. (2013). A Behavioral demonstration of overconfidence in judgment. *Psychological Science, 24*(7), 1190–1197. http://dx.doi.org/10.1177/0956797612470700.

Martins, L. L., Eddleston, K. A., & Veiga, J. F. (2002). Moderators of the relationship between work-family conflict and career satisfaction. *Academy of Management Journal, 45*, 399–409.

Matthews, R. A., & Barnes-Farrell, J. L. (2010). Development and initial evaluation of an enhanced measure of domain flexibility for the work and family domains. *Journal of Occupational Health Psychology, 15*, 330–346.

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. http://dx.doi.org/10.1037/0021-9010.79.4.599.

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

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. http://dx.doi.org/10.1016/0010-0285(81)90010-4.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Salgado, J. F. (2010). Cognitive abilities. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 255–275). New York, NY: Routledge.

Önkal, D., Yates, J. F., Simga-Mugan, C., & Öztin, Ş. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior and Human Decision Processes, 91*(2), 169–185. http://dx.doi.org/10.1016/S0749-5978(03)00058-X.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29*(3), 261–265. http://dx.doi.org/10.1037/h0022125.

Picone, P. M., Dagnino, G. B., & Minà, A. (2014). The origin of failure: A multidisciplinary appraisal of the hubris hypothesis and proposed research agenda. *Academy of Management Perspectives, 28*(4), 447–468. http://dx.doi.org/10.5465/amp.2012.0177.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864–901. http://dx.doi.org/10.1037/a0019737.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences, 23*(1), 125–133.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review, 120*(3), 697–719. http://dx.doi.org/10.1037/a0033152.

Revelle, W. (2014). psych: Procedures for personality and psychological research. R package version 1.4.2 Retrieved from <http://personality-project.org/r>.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40*(2), 193–218.

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review, 33*(2), 7–17.

Rynes, S. L., Brown, K. G., & Colbert, A. E. (2002). Seven common misconceptions about human resource practices: Research findings versus practitioner beliefs. *Academy of Management Executive, 16*(3), 92–103. http://dx.doi.org/10.5465/AME.2002.8540341.

Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes, 121*(2), 246–255.

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods or prediction. *American Journal of Sociology, 48*, 598–602.

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. http://dx.doi.org/10.1037/0033-2909.124.2.262.

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annual Review of Organizational Psychology and Organizational Behavior, 1*(1), 45–65. http://dx.doi.org/10.1146/annurev-orgpsych-031413-091255.

Seale, D. A., & Rapoport, A. (1997). Sequential decision making with relative ranks: An experimental investigation of the "secretary problem". *Organizational Behavior and Human Decision Processes, 69*(3), 221–236. http://dx.doi.org/10.1006/obhd.1997.2683.

Shariatmadari, D. (2015, July 18). Daniel Kahneman: "What would I eliminate if I had a magic wand? Overconfidence." *The Guardian*. <http://www.theguardian.com/books/2015/jul/18/daniel-kahneman-books-interview> Retrieved February 2, 2016.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Performance, 42*, 271–283.

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making, 18*(1), 29–53.

Siegel-Jacobs, K., & Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes, 65*(1), 1–17.

Slaughter, J. E., & Kausel, E. E. (2013). Employee selection decisions. In S. Highhouse, R. S. Dalal, & E. Salas (Eds.), *Judgment and decision making at work*. SIOP organizational frontiers series (pp. 57–79). New York, NY.

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). Outsmart your own biases. *Harvard Business Review, 93*, 65–72.

Stevens, C. K. (1998). Antecedents of interview Interactions, interviewers' ratings, and applicants' reactions. *Personnel Psychology, 51*(1), 55–85. http://dx.doi.org/10.1111/j.1744-6570.1998.tb00716.x.

Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In *Principles of forecasting: A handbook for researchers and practitioners* (pp. 106).

Swets, J., Dawes, R., & Monahan, J. (2000). Better decisions through science. *Scientific American, 283*, 82–87.

Tenbrunsel, A. E., & Diekmann, K. A. (2002). Job-decision inconsistencies involving social comparison information: The role of dominating alternatives. *Journal of Applied Psychology, 87*(6), 1149–1158. http://dx.doi.org/10.1037/0021-9010.87.6.1149.

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18*(1), 46–50.

Terpstra, D. E. (1996). The search for effective methods. *HR Focus, 73*(5), 16–17.

Terpstra, D. E., & Rozell, E. J. (1997). Why some potentially effective staffing practices are seldom used. *Public Personnel Management, 26*, 483–495.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. New York: Crown.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis – A useful supplement to regression analyses. *Journal of Business and Psychology, 26*, 1–9.

Tsai, C. I., Klayman, J., & Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. *Organizational Behavior and Human Decision Processes, 107*, 97–105.

Van der Zee, K. I., Bakker, A. B., & Bakker, P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology, 87*(1), 176.

Vredeveldt, A., & Sauer, J. D. (2015). Effects of eye-closure on confidence-accuracy relations in eyewitness testimony. *Journal of Applied Research in Memory and Cognition, 4*(1), 51–58. http://dx.doi.org/10.1016/j.jarmac.2014.12.006.

Whitecotton, S. M. (1996). The effects of experience and a decision aid on the slope, scatter, and bias of earnings forecasts. *Organizational Behavior and Human Decision Processes, 66*(1), 111–121. http://dx.doi.org/10.1006/obhd.1996.0042.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30*(1), 132–156. http://dx.doi.org/10.1016/0030-5073(82)90237-9.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall Inc..

Yates, J. F. (2010). Culture and probability judgment. *Social and Personality Psychology Compass, 4*(3), 174–188. http://dx.doi.org/10.1111/j.1751-9004.2009.00253.x.

Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes, 74*(2), 89–117. http://dx.doi.org/10.1006/obhd.1998.2771.

Yates, J. F., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: The "consumer's " perspective. *International Journal of Forecasting, 12*, 41–56.

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology, 33*(4), 345–366.