Using narratives and numbers in performance prediction

Niessen, A. Susan M.; Kausel, Edgar; Neumann, Marvin

Published in:
International Journal of Selection and Assessment

DOI:
10.1111/ijsa.12364

Publication date:
2022

Citation for published version (APA):
Niessen, A. S. M., Kausel, E., & Neumann, M. (2022). Using narratives and numbers in performance prediction: Attitudes, confidence and validity. *International Journal of Selection and Assessment*, *30*(2), 216-229. https://doi.org/10.1111/ijsa.12364

Received: 13 September 2021 | Revised: 21 November 2021 | Accepted: 24 November 2021

DOI: 10.1111/ijsa.12364

RESEARCH ARTICLE

International Journal of Selection and Assessment

WILEY

Using narratives and numbers in performance prediction: Attitudes, confidence, and validity

A. Susan M. Niessen[1] | Edgar E. Kausel[2] | Marvin Neumann[1]

[1]Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands

[2]School of Management, Pontificia Universidad Católica de Chile, Santiago, Chile

Correspondence
A. Susan M. Niessen, Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.
Email: a.s.m.niessen@rug.nl

Abstract

In a preregistered prediction-task experiment, we investigated the effect of narrative versus quantified information on decision-maker perceptions, confidence, predictor weighting, and predictive accuracy when making performance predictions. We also investigated the effect of who quantifies information (the decision maker or someone else). As expected, we found higher perceived informativeness and use intentions for narrative than quantified information. Information presented narratively was also weighted somewhat more heavily than quantified information. Using quantitative information quantified by decision makers themselves yielded higher perceived autonomy and use intentions than quantitative information quantified by someone else. However, no differences in prediction confidence were found and self- and other-produced quantifications received identical weight. Moreover, unexpectedly, differences in weighting did not translate to differences in predictive accuracy.

KEYWORDS

narratives, nonvalid information, performance prediction, quantification, validity

Practitioner points

- Several authors suggested that narrative information is perceived as richer, and hence, is more influential in judgments and predictions, than quantitative information.
- Additionally, quantified information was suggested to be more influential when the decision-makers quantify information themselves.
- We found that narrative information was perceived as more informative, yielded higher use intentions, and was weighted somewhat more heavily than quantified information. We also found higher perceived autonomy and use intentions for self-quantified, than for other-quantified information, but no differences in assigned weight.
- Notably, we found no differences in confidence or predictive accuracy.
- Since we did not find the expected effects on predictive accuracy, the implications for selection practice remain unclear.

1 | INTRODUCTION

Predictions of future performance and behavior are the core of hiring and admission decisions. Usually, these predictions are made using several sources of information such as test scores, resumes, cover letters, and interviews. Traditionally, the field of performance prediction has focused on investigating relations between quantitative or quantified ratings derived from assessment instruments and future performance (Van Iddekinge & Ployhart, 2008), typically using correlations and multiple regression analyses. However, information derived from some commonly used instruments, such as cover letters and interviews, is often not formally quantified in practice (Highhouse, 2008), but observed and interpreted in qualitative or narrative form. Moreover, even when information derived from assessment instruments is quantified, optimal statistical models are usually not used to make decisions in practice (Kuncel, 2008; Ryan & Sackett, 1987; Slaughter & Kausel, 2013). Instead, information is combined holistically, without applying weights explicitly, optimally, or consistently (Dawes, 1979; Highhouse & Kostek, 2013; Karelaia & Hogarth, 2008; Yu, 2018). Therefore, typical validity studies may not accurately reflect the validity of these types of instruments when used in practice (Kuncel et al., 2013; Slaughter & Kausel, 2013).

2 | NARRATIVES AND NUMBERS

Information can be presented quantitatively (as numeric scores or ratings, such as test scores or interview ratings) or qualitatively (in behavior, words or images, such as responses to interview questions or written personal statements). Narrative information is qualitative information in the form of a story about oneself (first-person; Beach, 2010) or about someone else (third-person; Winterbottom et al., 2008).

It has been suggested that information presented in narrative form is perceived as richer in information than quantitative information (Kuncel, 2018). This could explain the generally high face validity of instruments with narrative characteristics (e.g., interviews), compared to instruments that generally yield quantitative outcomes (e.g., cognitive ability tests, Anderson et al., 2010; Niessen et al., 2017). Quantification is seen as dehumanizing (Dawes, 1979) and reductionistic, removing relevant information that reflects the context and the uniqueness of an individual (Boswell et al., 2003; Longoni et al., 2019; Meehl, 1954; Newman et al., 2020). Therefore, we expect that narrative information is perceived as more informative for making predictions of future performance than quantified information. Moreover, this perceived richness elicits "sense-making" (Dana et al., 2013; Pennington & Hastie, 1992); narratives allow the construction of a coherent story about a person, which makes human judges more confident in their predictions about those persons (Kahneman, 2011; Slaughter & Kausel, 2013). Consequently, we expect that decision makers are more inclined to use narrative information than quantified information, both in terms of use intentions in prediction procedures, and in terms of weight given to information when making holistic predictions (Kuncel, 2018). When we refer to "weight" in holistic predictions, we mean weight from a Brunswikian Lens model perspective (Brunswik, 1956): the relative importance or impact of certain information on performance predictions.

Research on patient decision-making (Winterbottom et al., 2008) and the acceptance of general conclusions (Allen & Preiss, 1997) showed mixed findings on the effect of narrative versus quantitative information. However, when it comes to making individual decisions rather than conclusions about general principles, it seems that information presented in narrative form was more influential than quantitative information, especially when narratives were written in the first-person (Winterbottom et al., 2008). Such first-person narratives are common in performance prediction, for example in hiring interviews, cover letters, and personal statements. Moreover, in a study on predicting finishing times of marathon runners, Sanfey and Hastie (1998) found differences in how information was weighted depending on presentation format of all provided information (textual vs. tabular or graphical), demonstrating that these format differences can indeed affect how information is weighted.

However, although it has been suggested and theorized (Kuncel, 2018, p. 476), we were unable to identify empirical evidence that supports overweighting of narrative information in the context of performance prediction. Therefore, the first aim of this study was to investigate if presenting information in narrative or quantified form affects decision makers' perceptions of the information presented, their confidence, and the way they use information. We have the following expectations:

Information presented in narrative form will (a) be perceived as more informative, (b) result in higher prediction confidence, (c) yield higher use intentions, and (d) will be weighted more heavily, than information presented in quantified form when making holistic predictions.

3 | QUANTIFICATION AND "QUANTIFIERS"

If using information in quantitative form indeed has benefits, narrative indicators of performance would have to be quantified first. It has also been argued that the influence of quantified information depends on who does the quantifying, and hence, processes the narrative information (Kausel et al., 2016). Kahneman (2011, p. 225) asserted that quantified information is more influential when the decision maker is the one who quantifies narrative information. For this reason, Kuncel (2018, p. 476) recommends that those involved in data combination should not be involved in data collection (e.g., quantifying), to mitigate overweighting quantified narrative information that is perceived as having a "rich" basis. However, we were unable to find direct empirical studies that provide support for this recommendation. Therefore, a second aim of this study was to investigate whether the person who quantifies narrative information (the decision maker or someone else), affects decision-maker perceptions and how they weight information.

The recommended separation of data collection and data combination likely hurts the perceived autonomy of decision makers (i.e., perceived control over an outcome), because the decision maker has no control over how the narrative information was processed and quantified (Dalal & Bonaccio, 2010). Perceived autonomy is considered a central human need (Deci & Ryan, 1991), and prior studies have found that it was related to use intentions of performance prediction procedures (Dipboye & Jackson, 1999; Dipboye, 1997; Nolan & Highhouse, 2014). The advice-taking literature suggests that reduced autonomy, and thus, control over the judgment process could result in lower use intentions. This occurs because of a lack of insight in how another person interpreted the information when quantifying it, which reduces confidence in the validity of the quantified information (Yaniv & Kleinberger, 2000; Yaniv, 2004). Indeed, several studies show that people have more confidence in, and rely more on, their own judgment than the judgments of others (Bonaccio & Dalal, 2006; Yaniv, 2004). Consequently, we expect that decision makers are more inclined to use quantified information when they quantified it themselves, both in terms of use intentions for prediction procedures, and in terms of weight given to information when making predictions. We have the following hypotheses:

Self-produced quantifications will yield (a) more perceived autonomy, (b) higher prediction confidence, (c) higher use intentions, and (d) will be weighted more heavily, than other-produced quantifications when making holistic predictions.

4 | NONVALID INFORMATION AND PREDICTIVE ACCURACY

Decision makers concerned with performance prediction usually have access to both valid and nonvalid information regarding the performance prediction at hand (e.g., Highhouse, 1997). This is not problematic when information is combined based on mechanical, data-based methods such as regression models, which is the default approach in academic research on the predictive validity of predictors and procedures. Regression models are excellent at ignoring nonvalid information by assigning small weights, so the only disadvantage is inefficiency caused by collecting redundant information. In contrast, access to nonvalid information is problematic when information is combined holistically, because human decision makers are unable to ignore nonvalid information (Kemmelmeier, 2004). This results in overweighting non- or less valid information, which "dilutes"[1] the valid information and hurts predictive accuracy (Hall et al., 2007; Nisbett et al., 1981). From a Brunswikian Lens model perspective (Brunswik, 1956), this would result in a mismatch between how decision makers weight the information (subjects model) versus how the information should be weighted to yield accurate predictions (ecological model), resulting in low validity of decision makers' predictions.

Importantly, the most valid indicators of performance, such as scores on standardized ability tests and assessment center ratings, tend to be quantitative in nature, while indicators that are typically observed and used in narrative form, such as unstructured hiring interviews, cover letters, and personal statements, tend to have substantially lower validity (Sackett et al., 2017; Schmidt & Hunter, 1998). Therefore, if narrative information indeed has more influence on judgments and decisions, as hypothesized above, nonvalid information presented in narrative form would be expected to "dilute" other, more valid information to a larger extent than quantitative information. Similarly, if information quantified by decision makers themselves receives more weight than information quantified by someone else, larger dilution effects would be expected for self-quantified than for other-quantified information.

Two studies (Dana et al., 2013; Kausel et al., 2016) directly investigated the effect of adding nonvalid information to valid information on predictive accuracy in the context of human performance prediction, both using the unstructured interview as the less valid (e.g., Huffcutt et al., 2014) additional predictor. In Dana et al. (2013), decision makers observed interviews in narrative form; while in Kausel et al. (2016) they received information on the interview as quantitative ratings made by someone else. Both found evidence of reduced predictive accuracy, and Kausel et al. (2016) demonstrated that reduced predictive accuracy was related to overweighting the interview ratings and underweighting valid information (dilution). Thus, both studies demonstrated that predictive accuracy suffered as a result of adding less valid information to valid information, when that additional information was presented in narrative form or numeric form quantified by others. Nevertheless, little research has been conducted on factors that influence the weighting of information in performance predictions, despite calls for more studies on decision making and information utilization on performance predictions (Kuncel, 2018; Neumann et al., 2021). Furthermore, these studies do not allow for a comparison in terms of the extent of negative effects on predictive accuracy, because of the different designs and analytical approaches used. Therefore, we investigate the differences depending on information format (narrative vs. quantified) and quantification source (decision maker vs. someone else) on predictive accuracy. We have the following hypotheses:

When less valid information is presented in addition to valid information…

(a) Presenting information in narrative form will result in lower predictive accuracy, compared to presenting information in quantified form.

(b) Using self-produced quantifications will result in lower predictive accuracy, compared to other-produced quantifications.

5 | PRESENT STUDY

The preregistration protocol is available on OSF. Personal statements of applicants to an undergraduate psychology program were used as the focal stimulus material. To investigate the effect of information format (narrative vs. quantified) and quantification source (decision maker vs. someone else) on decision-maker perceptions, information weighting, and accuracy, we used a within-subjects design in which participants predicted academic success of 10 applicants, using their personal statements in the form of (1) narratives, (2) self-produced quantifications, and (3) quantifications produced by someone else (other-produced quantifications). Additionally, a planned exploratory mediation analysis is conducted to investigate relations between information format and quantification source, and perceived informativeness, autonomy, confidence, and use intentions.

It is not possible to use self-produced quantifications to make predictions without first being exposed to the personal statements in narrative form. To ensure that the self-produced quantifications (and not the statements in narrative form) would be the most salient basis of participants' predictions, we asked participants to provide ratings of 10 personal statements about one week before participating in the prediction tasks, assuming that this time lag is long enough to prevent remembering the exact narrative content of the statements. Those ratings were presented to participants as the self-produced quantifications in the focal part of the experiment.

To be able to investigate information weighting, other predictors than the personal statements had to be included in the prediction task. We chose to include high school GPA and an admission test score. These are valid predictors of academic success, allowing for an investigation of dilution, and commonly used, making the prediction task representative for selection decisions in practice.

6 | METHOD

6.1 | Participants

Based on a power analysis (see the preregistration), we planned to collect responses of at least $n = 65$ usable participants, so the responses to the attention checks were inspected during data collection. Because the number of required usable participants was not met before the predetermined deadline (see the preregistration protocol), we extended the data collection by one week to meet the minimum number of participants. In total, 87 participants completed both parts of the study, of whom 19 did not pass at least one of the attention checks in one of the parts of the study (see Supporting Information S1 for details). These participants were excluded from the data set. The final sample size was $n = 68$. A sensitivity analysis showed that, as intended, this resulted in small to moderate detectable minimum effect sizes with statistical significance ($n^2_p = 0.05$ for omnibus tests and $d = 0.43$ for contrasts, when correlations between the measures are conservatively set to zero). Since all participants made 10 predictions in each condition, $k = 680$ predictions per condition were collected. All participants were 1st year Psychology students and had the Dutch nationality. Their mean age was 19 ($SD = 1.73$), and 93% was female.

6.2 | Stimulus material and criterion variables

The material used in the prediction task included high school GPA, scores on an admission test, and the personal statement of 192 students who applied to and were accepted for the psychology undergraduate program in 2014. High school GPA was the mean grade obtained in high school and the admission test was an exam about introductory psychology material. Both were shown using the Dutch grading scale (1–10, with 10 being the highest grade) and were good predictors of academic achievement (Niessen et al., 2018). The personal statement was about 250 words in length and was meant to demonstrate motivation for the program. Identifiable information such as names and addresses were omitted to ensure anonymity of the applicants. In addition, ratings of personal statements made by the participants approximately one week before performing the prediction tasks were used as the self-produced quantifications, and ratings of the personal statements made by participants in a prior study among similar participants, using the same question, were used as the other-produced quantifications.

First year GPA and dropout were obtained from the university administration to evaluate predictive accuracy. First year GPA was the mean grade across all 1st year courses, and dropout was recorded as a binary variable.

6.3 | Procedure and design

Participants were informed that the study consists of two parts and signed up for both parts at the same time. In part I, participants were asked to rate personal statements of 10 applicants to a psychology undergraduate program. This task was conducted online. The applicants were distributed randomly among participants. Before providing ratings, they were informed that they would use these ratings again in part II of the study to make predictions about the applicants' academic performance. They also learned that they could earn up to 5 euros depending on their performance in part II. Attentive responding was checked using two attention checks (see the preregistration protocol). The median time between participating in part I and part II was 6 days.

Part II of the study was conducted in a lab space at the university and utilized a within-subjects design with three conditions. Participants were informed that they would make predictions of 10 applicants' academic performance in each condition, and (again) that they could earn up to 5 euros depending on their predictive performance in the study (see the OSF project page for examples of the prediction task and the reward scheme). In each condition, participants were presented with the applicants' high school GPA, admission test score, and personal statement information that varied in format across conditions. Participants were told that high school GPA and the admission test score were good predictors of academic performance, and that the personal statement was a poor predictor of academic performance. They were asked to predict each applicant's 1st year GPA, and whether the applicant would drop out in the 1st year. They also indicated how confident they were about each prediction. To provide a common frame of reference, mean, minimum and maximum scores for high school GPA and the admission test scores were presented on the screen, as was the mean, minimum, and maximum 1st year GPA and the percentage of students who drop out in the 1st year.

In the *narrative condition*, participants were presented with each applicant's personal statement. They were first asked to rate it (without seeing the admission test score and high school GPA), using the same scale as in part I, before making their predictions. In the *self-produced quantification condition*, participants were presented with the same applicants whose personal statements they rated in part I (when they only saw and rated the personal statement). They were presented with the rating they provided in part I, but not with an applicant's personal statement in narrative form. In the *other-produced quantification condition*, they were presented with ratings provided by others. The order of conditions was randomized and participants saw unique applicants in each condition.

After making all predictions in a single condition, participants completed questions about their perceived autonomy, the informativeness, and use intentions of the information they could use to make their predictions. After completing all conditions, they ranked the three types of information in terms of preference to use for future admission decisions. The median time it took to complete the study was 13 min for part I and 42 min for part II. Attentive responding for part II was checked using two attention checks after general instructions and one attention check after presenting the instructions in each condition (see the preregistration protocol for details). All materials were presented in Dutch.

6.4 | Measures

6.4.1 | Personal statement ratings

In part I and in the narrative condition in part II, participants rated the personal statements of each applicant they judged based on the following question: *How do you rate the motivation of this applicant to study psychology at the University of Groningen?* Ratings were provided on a seven-point scale, from *not motivated at all* to *very motivated*.

6.4.2 | Predicted academic performance

For each applicant, participants were asked: *Based on the information provided, what do you think this applicant's first year GPA will be?* The response was given on the Dutch grading scale (1 to 10, with increments of 1 decimal point). They were also asked: *Based on the information provided, do you think this applicant will drop out in the first year?* (*Yes*, *No*).

6.4.3 | Prediction confidence

Participants were asked to indicate their confidence in each prediction they made. For this purpose, the GPA predictions were dichotomized as being higher or lower than 6. This grade was chosen because it is the minimum passing grade (69% of the applicants had a GPA of 6 or higher). Depending on their prediction, participants answered the following question: *You predicted a GPA of (6 or higher/lower than 6) for this applicant. How confident are you that this applicant will obtain a GPA of (6 or higher/lower than 6)?* Depending on their prediction of dropout, they also answered: *You predicted that this applicant will/will not drop out in the first year. How confident are you that this applicant will/will not drop out?* Both questions were answered using a slider ranging from 50% to 100%.

6.4.4 | Informativeness

After each condition, participants indicated their perceived informativeness of the information they could use, using a two-item scale based on Dana et al. (2013): *I am able to infer a lot about this person given the information provided* and *I got information that was valuable in making predictions*. Responses were provided on a five-point scale (*strongly disagree—strongly agree*). Across conditions, the scale yielded a reliability of $\overline{\alpha}$ = .63.

6.4.5 | Perceived autonomy

After each condition, participants indicated how much autonomy they perceived when making their predictions using a six-item scale adapted from Nolan and Highhouse (2014). An example item is: *The information I could use to make my predictions gave me a sense of control*. Responses were provided on a five-point scale (*strongly disagree—strongly agree*). Across conditions, the scale yielded a reliability of $\overline{\alpha}$ = .72.

6.4.6 | Use intentions

After each condition, participants indicated if they would use the information they just used to make future admission decisions, using a three-item scale based on Nolan and Highhouse (2014). An example item is *I would choose to use this information to make future admissions decisions*. Responses were provided on a five-point scale (*strongly disagree—strongly agree*). Across conditions, the scale yielded a reliability of $\overline{\alpha}$ = .74. As a measure of use intentions based on joint evaluation (Hsee et al., 1999) participants also ranked the information they received in the three conditions in order of preference for use in future admission decisions, after completing all conditions.

6.5 | Analyses

Means of respondents' confidence were computed across the 10 cases judged in each condition. These means were used in repeated measures (RM) ANOVAs, and planned contrasts (H1b, 2b) were conducted when a significant difference was found in the omnibus test. RM-ANOVAs and planned contrasts were also used to investigate differences in informativeness, autonomy, and use intentions (H1a, 2a, 1c, and 2c). Differences in preference rankings (H1c, 2c) were analyzed using a Friedman test, and follow-up tests were conducted using Wilcoxon's signed-rank tests.

To investigate if information format and quantification source (H1d, 2d) affected the weight assigned to the personal statements, the weighting policies of participants were analyzed by creating judgmental bootstrapping models (Armstrong, 2001, also referred to as "models of man," e.g., Goldberg, 1970) using multiple and logistic regression and relative weights analysis. One model per condition was created using all 680 predictions made by participants as the dependent variable and the three predictors as independent variables. For the relative weights analyses we used the *relaimpo* package in R (Grömping, 2006) for GPA predictions and the code for logistic regression models provided by Tonidandel and LeBreton (2015) for dropout predictions. Differences were not formally tested for statistical significance, but were interpreted based on differences in relative weights.

To test differences in predictive accuracy (H3a, 3b), correlations were computed between predicted GPA and actual GPA, and predicted dropout and actual dropout, in each condition. The differences were tested for statistical significance using the test for dependent correlations with different variables (Steiger, 1980) using the *psych R* package (Revelle, 2019). To compare participants' predictive accuracy to optimal predictions, multiple (logistic) regression analyses with the three predictors regressed on the actual outcomes were performed for each condition.

Finally, we explored relations between information format and quantification source and perceived informativeness, autonomy, confidence, and use intentions based on correlations and mediation analysis using the MEMORE macro for mediation and moderation analysis in within-subjects designs (Montoya & Hayes, 2017).

7 | RESULTS

Table 1 shows that the randomized distribution of stimulus cases resulted in very similar distributions of scores and ratings and similar predictive validities for the criterion measures. High school GPA was a good predictor of both criteria, with $r \approx .47$ for GPA and $r \approx -.24$ for dropout. The admission test score was also a good predictor, with $r \approx .41$ for GPA and $r \approx -.28$ for dropout. The quantitative ratings of the personal statements had more variable but consistently low to near-zero predictive validity, with $r \approx .06$ for GPA and $r \approx .03$ for dropout.

7.1 | Perceptions[2]

Descriptive statistics for perceived informativeness, autonomy, confidence, and use intentions are shown is Supporting Information S2. Checking the assumptions for RM-ANOVA showed no serious violations of the normality assumption. For the informativeness and autonomy scales, one substantial outlier was detected (extreme standardized score and residual accompanied by a high Cook's distance in one of the conditions). All analyses were repeated with the outlier excluded and the results were virtually identical. Therefore, results on the full sample are reported (results with the outlier removed are available on the OSF project page).

TABLE 1 Descriptive statistics and predictive validity of the stimulus material in each condition

	Condition		
Variable	Narrative	Self-produced quantifications	Other-produced quantifications
High school GPA			
M	6.61	6.63	6.63
SD	0.45	0.45	0.48
r GPA	0.49	0.45	0.47
r dropout	−0.24	−0.24	−0.22
Admission test score			
M	6.31	6.49	6.48
SD	1.77	1.79	1.79
r GPA	0.41	0.41	0.42
r dropout	−0.29	−0.28	−0.28
Personal statement rating			
M	5.35	5.36	5.57
SD	1.14	1.38	1.08
r GPA	0.11	0.03	0.04
r dropout	−0.04	0.05	0.08
First year GPA			
M	6.35	6.29	6.30
SD	1.09	1.24	1.17
Proportion dropped out	0.11	0.13	0.13

7.1.1 | Informativeness

Figure 1 shows that personal statements in narrative form were perceived as most informative and quantifications provided by others were perceived as least informative. The omnibus test of an RM-ANOVA indeed showed significant differences in perceived informativeness between the conditions ($F_{(2, 134)} = 18.81$, $p < .001$, $n^2_p = 0.22$). A planned contrast comparing the narrative personal statements to both quantified forms showed a moderate difference ($t_{(134)} = 5.48$, $p < .001$, $d = 0.52$), providing support for H1a.

7.1.2 | Autonomy

Figure 1 shows that participants perceived most autonomy when they used personal statements in narrative form, and least autonomy when they used quantifications provided by others. Significant differences in perceived autonomy between the conditions were found ($F_{(2, 134)} = 25.25$, $p < .001$, $n^2_p = 0.27$). A planned contrast showed a moderate difference in autonomy perceptions between the self-produced quantifications and the other-produced quantifications ($t_{(134)} = 5.13$, $p < .001$, $d = 0.48$), providing support for H2a.

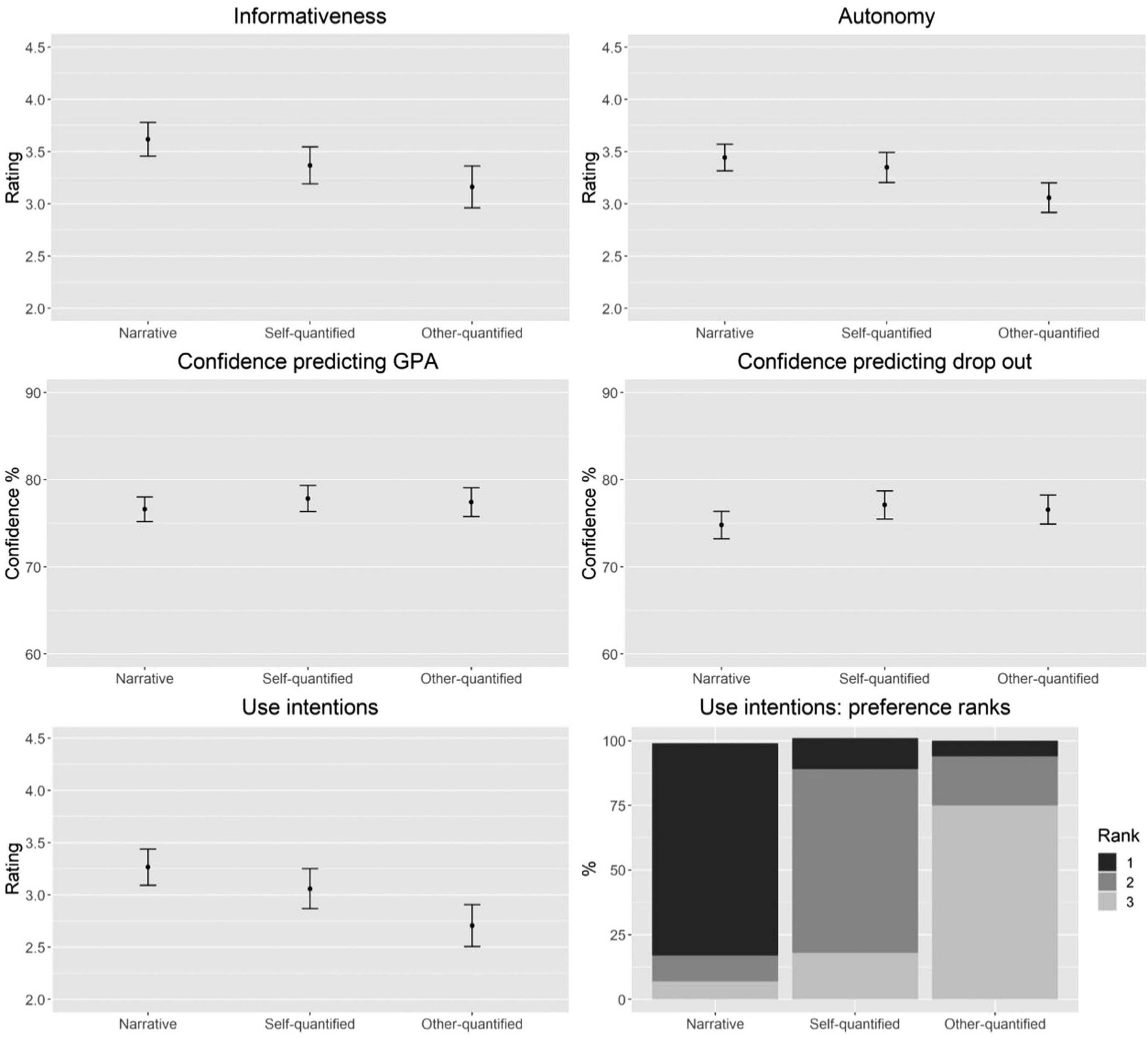

FIGURE 1 Perceived informativeness, autonomy, confidence, and use intentions. The *y* axes ranges were set at approximately 1.5 SD from the means (Witt, 2019), with results rounded up for upper bounds and down for lower bounds, and assuring that all variables measured on the same scale have the same *y* axis range. Error bars represent 95% CIs

7.1.3 | Confidence

Participants reported the highest confidence when they used their own quantifications for both criterion measures (see Figure 1). For predicting GPA, the omnibus test showed no significant differences in confidence between conditions ($F_{(2, 134)} = 1.72$, $p = .18$, $n^2_p = 0.03$), so no planned contrast tests were conducted. For predicting dropout, the omnibus test did show significant differences in confidence between the conditions ($F_{(2, 134)} = 5.03$, $p = .01$, $n^2_p = 0.07$). Small differences in confidence were found between using narrative personal statements and quantifications ($t_{(134)} = -3.09$, $p = .003$, $d = -0.32$), but in the opposite direction as expected. No statistically significant differences in confidence between predictions made using the self-produced quantifications or other-produced quantifications were found ($t_{(134)} = 0.72$, $p = .47$, $d = 0.08$). Therefore, H1b and H2b were not supported.

7.2 | Use intentions

Figure 1 shows that participants reported the highest use intentions when they used personal statements in narrative form, and lowest use intentions when they used quantifications provided by others. Mauchly's test indicated a violation of the sphericity assumption ($\chi^2_{(2)} = 0.88$, $p = .02$), so the Greenhouse-Geiser correction was applied. The omnibus test showed significant differences in use intentions between the conditions ($F_{(1.79, 119.95)} = 16.53$, $p < .001$,

$n^2_p = .20$). Moderate differences in use intentions were found between narrative and quantified information ($t_{(134)} = 4.49$, $p < .001$, $d = 0.53$), and between self-produced quantifications and other-produced quantifications ($t_{(134)} = 3.59$, $p < .001$, $d = 0.43$), both in the expected direction.

Figure 1 also shows the results for the joint-evaluation preferences rankings. Most participants preferred to use personal statements in narrative form for future decisions and ranked using quantifications made by others last. A Friedman test showed significant differences in mean ranks between conditions ($\chi^2_{(2)} = 70.97$, $p < .001$, $w = .52$). Follow-up Wilcoxon's signed ranks test showed a moderate difference between using narrative personal statements and self-produced quantifications ($z = 5.19$, $p < .001$, $r = .45$), a large difference between narrative personal statements and other-produced quantifications ($z = 6.62$, $p < .001$, $r = .57$), and a moderate difference between the self-produced quantifications and the other-produced quantifications ($z = 4.52$, $p < .001$, $r = .39$). The findings based on separate and joint evaluation both provide support for H1c and H2c.

7.3 | Correlations and exploratory mediation models

In addition to testing the hypotheses, we explored relations between perceptions of informativeness and autonomy, confidence, and use intentions. Generally, we would expect positive relations between all variables. All correlations are presented in Table 2. No substantial differences in correlations were observed between conditions (see Supporting Information S3), so they were averaged across conditions. Remarkably, confidence showed no relation with informativeness or autonomy. Also, we found no or a small negative relation between confidence and use intentions.

We also explored mediation. We intended to explore the serial indirect effect of information format (narrative or quantified) on use intentions, through perceived informativeness, via confidence, and the serial indirect effect of quantification source (decision maker vs. someone else) on use intentions, through perceived autonomy, via confidence. However, since we did not find the expected relations between confidence and any of the other variables, these models were not likely to show good fit. Therefore, we dropped confidence as a mediator in our models. Instead, we explored a model in which the effect of information format on use intentions was mediated by informativeness, and a model in which the effect of quantification source on use intentions was mediated by autonomy.

As demonstrated in Figure 2, the first model suggests that information format (narrative or quantified) had a direct effect on perceived informativeness, which had a direct effect on use intentions. The indirect effect accounted for 40% of the total effect (see Preacher & Kelley, 2011) and the bootstrapped 95% confidence interval around the indirect effect coefficient does not include zero (see Supporting Information S4). This suggests that perceived informativeness partially mediates the effect of information format on use intentions; narrative information is perceived as more informative, which in turn, is related to higher use intentions.

Furthermore, the second model suggests that quantification source (decision maker vs. someone else) had a direct effect on perceived autonomy, which in turn had a direct effect on use intentions (see Figure 3). Forty-two percent of the total effect was accounted for by this indirect effect and its 95% confidence interval excluded zero (see Supporting Information S5), suggesting at least partial mediation of the relation between quantification source and use intentions through perceived autonomy.

7.4 | Judgmental bootstrapping models and relative weights

Linear models regressing participants' prediction on the three predictor variables showed very good fit. For predicted GPA, these models resulted in $R = .89$ in the narrative condition, $R = .90$ in the self-produced quantification condition, and $R = 0.90$ in the other-produced quantification condition. For predicted dropout, the models resulted in pseudo R^2's of $R_p^2 = 0.55$ in the narrative condition, $R_p^2 = 0.51$ in the self-produced quantification condition, and $R_p^2 = 0.42$ in the other-produced quantification condition. These results show that participants weighted the predictors quite consistently, with some but small differences between conditions for predicting dropout.

Figure 4 shows the relative weights as assigned to each predictor by participants in each condition, and the relative weights as assigned based on an optimal prediction model. These results support the expectation that narrative information would be weighted more heavily compared to information presented in

TABLE 2 Correlations between perceptions of informativeness and autonomy, confidence, and use intentions

Variable	Informativeness	Autonomy	Confidence (GPA)	Confidence (dropout)
Autonomy	0.35			
Confidence (GPA)	−0.06	−0.02		
Confidence (dropout)	−0.07	−0.03	0.78	
Use intentions	0.63	0.18	−0.04	−0.14

Note: Averaged across conditions using Fisher's *z* transformation.

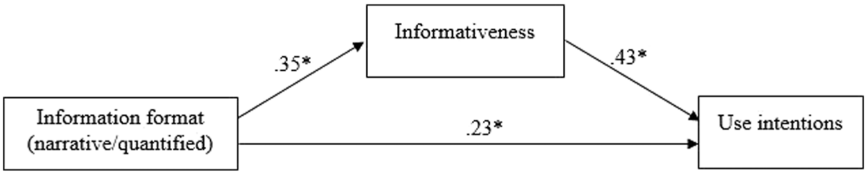

FIGURE 2 Mediation model for the indirect effect of information format on use intentions through informativeness

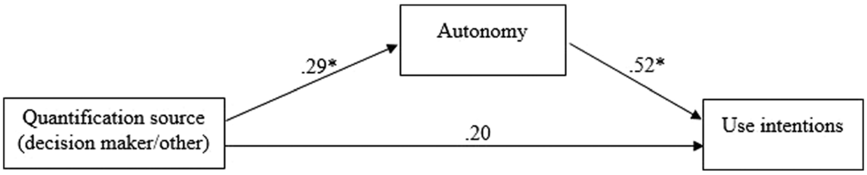

FIGURE 3 Mediation model for the indirect effect of quantification source on use intentions through autonomy

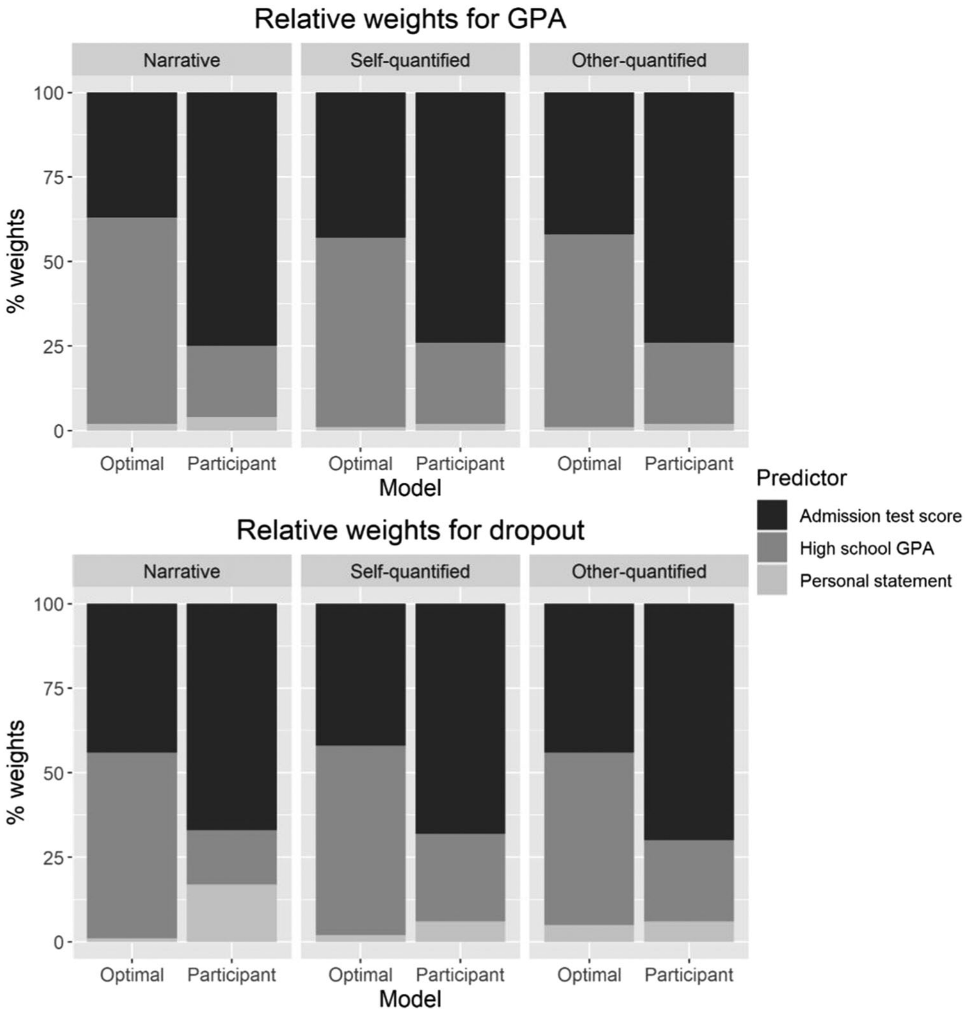

FIGURE 4 Relative weights assigned to the predictors in each condition, by the participants and based on optimal regression models

quantified form (H1d), most notably when predicting dropout. For predicting GPA, the personal statements received little weight in all conditions, but were weighted twice as heavily when presented in narrative form (4%), compared to quantified form (2% in both quantification conditions). For predicting dropout, the personal statement was weighted more heavily and the difference was larger, with 17% in the narrative condition and 6% in both quantification conditions. However, the relative weights were identical in the self-produced and other-produced quantification conditions for predicting GPA and dropout, providing no support for the expectation that self-produced quantifications would be weighted more heavily than other-produced quantifications (H2d).

Figure 4 also shows that, compared to optimal weights, participants overweighted the personal statements, most notably when predicting dropout based on narrative and self-produced quantitative information. Participants also overweighted the admission test score and underweighted high school GPA. However, the latter is unlikely to affect predictive accuracy substantially due to their comparable predictive validities and high intercorrelation.

7.5 | Predictive accuracy

Correlations between the predictions of academic success made in each condition and the actual outcome measures are shown in Tables 3 and 4. When checking the assumptions, several outliers with extreme standardized residuals were detected in each condition for the GPA predictions. Therefore, the analyses were conducted with and without outliers. More details on the analyses with outliers excluded are in Supporting Information S6.

For GPA, the correlations were higher when participants were presented with quantifications of the personal statements, compared to seeing the statements in narrative form. The differences in correlations between the narrative condition and the self-produced quantification condition ($z = -1.11$, $p = .13$, $r_{\text{diff.}} = -.05$) or the other-produced quantification condition ($z = -1.22$, $p = .11$, $r_{\text{diff.}} = -.06$) were not statistically significant based on the entire data set. However, repeating the analyses without outliers resulted in larger and statistically significant differences ($z = -2.52$, $p = .01$, $r_{\text{diff.}} = -.11$ and $z = -2.05$, $p = .02$, $r_{\text{diff.}} = -.09$, respectively). For predicting dropout (Table 4), the differences in correlations between the narrative condition and the self-produced quantification condition ($z = 0.52$, $p = .30$, $r_{\text{diff.}} = .03$) and the other-produced quantification condition ($z = -0.02$, $p = .49$, $r_{\text{diff.}} < -.01$) were not statistically significant and very small or not in the expected direction. Therefore, most of these findings do not support the hypothesis that providing information in narrative form results in lower predictive accuracy compared to providing information in quantified form (H3a).

Similarly, the differences between the correlations obtained in the self-produced and the other-produced quantification conditions were very small and not statistically significant in the complete sample ($z = -0.12$, $p = .45$, $r_{\text{diff.}} = -.01$ for GPA, and $z = -0.54$, $p = .30$, $r_{\text{diff.}} = -.03$ for dropout), although they were in the expected direction. Results were similar when outliers were discarded (see Supporting Information S6). These findings do not provide support for the hypothesis that predictions made using quantifications provided by others would result in higher predictive validity than predictions made using self-produced quantifications (H3b).

Comparing the predictive accuracy of participants' predictions to optimal model predictions (Tables 3 and 4) shows that predictions made by participants were substantially less accurate than those based on regression models. Furthermore, the predictability based on an optimal regression model varied somewhat between conditions, due to random sampling of the stimulus cases. When using the complete data set, the narrative condition had the highest predictability based on an optimal linear model. These differences in predictability could be an explanation for the small differences detected between predictive accuracy in the different conditions. Comparisons between predictive accuracy based on regression models and participants' predictive accuracy, using the complete sample and when outliers are removed, do suggest that the difference in predictive

TABLE 3 Predictive accuracy obtained in each condition for predicting GPA

	Complete data set			Outliers removed		
	R		*R* difference:	*R*		*R* difference:
Condition	Participants	Optimal model	optimal—participant model	Participants	Optimal model	optimal—participant model
Narrative	0.38	0.53	0.15	0.43	0.61	0.18
Self-produced quantifications	0.43	0.49	0.06	0.54	0.64	0.10
Other-produced quantifications	0.44	0.51	0.07	0.52	0.66	0.14

TABLE 4 Predictive accuracy obtained in each condition for predicting dropout

	R	Pseudo R^2		Pseudo R^2 difference: optimal—
Condition	Participants	Participants	Optimal model	participant model
Narrative	0.22	0.05	0.15	0.10
Self-produced quantifications	0.19	0.04	0.13	0.09
Other-produced quantifications	0.22	0.05	0.12	0.07

accuracy between optimal models and participants' predictions were largest when statements were presented in narrative form.

7.6 | Additional analyses

In addition to correlational analyses, we also analyzed predictive accuracy by computing the mean absolute percentage errors (MAPE) for the GPA predictions and the number of correct dropout predictions in each condition. All differences between conditions were small and not statistically significant (see Supporting Information S7).

8 | DISCUSSION

The aim of this study was to investigate if the nature (narrative or quantified) and source of the quantification (the decision maker or someone else) affected the weights assigned to information in performance predictions, predictive accuracy, and decision-maker perceptions and attitudes. In line with expectations (Kahneman, 2011; Kuncel, 2018), participants indeed perceived information as more informative when personal statements were presented in narrative form than when quantitative ratings based on the same information were presented, and reported higher use intentions as a result. Participants also weighted the personal statements higher in making predictions when they were presented in narrative form.

However, surprisingly, we found little evidence that these higher weights attenuated predictive accuracy; we found no evidence for this when predicting dropout, and for predicting GPA we only found the expected differences when outliers were excluded, but not in the entire sample. When results differ depending on outlier exclusion, deciding what results to trust more is difficult. Because the outlying data were apparently not erroneous, we are reluctant to mainly rely on the data with outliers excluded and believe that interpreting the results as inconclusive is most appropriate.

A possible explanation for the absence of clear effects on predictive accuracy could be that, even when the weight assigned to the personal statements was higher when they were presented in narrative form, it was still quite low, especially for predictions of GPA. This could be because asking participants to rate the statements in the narrative condition before making their predictions made the task, or participants' approach to the task, more structured than would be the case without being asked to rate the statement first. This rating step could also have changed the way participants perceived the narrative information, even though they were presented with the personal statement in narrative form when making their predictions.

Another possible explanation is the information provided to participants, stating that the personal statement was a poor predictor, and the other two predictors were good predictors of academic achievement. Providing participants with this information resulted in a stringent test of the hypotheses, and a prior study found that providing educational information can improve predictive accuracy (Neumann et al., 2021). Providing predictor information could have had similar effects. In practice, the predictive value of information is likely less salient to decision makers (or validity beliefs are even incorrect, e.g. Lievens & De Paepe, 2004; Rynes et al., 2002). We also did not find that participants reported higher confidence in their predictions when they had access to a narrative personal statement, but even found a small effect in the opposite direction. This last finding seems at odds with the notion of sensemaking; that narratives allow decision makers to construct coherent stories, which should increase confidence (Dana et al., 2013). Perhaps informing participants that personal statements have poor validity influenced their behavior more than we anticipated.

Another possible explanation is that not only the weight assigned to the predictors varied between conditions, but also how consistently participants did that. Consistency in weighting is positively related to prediction accuracy (Karelaia & Hogarth, 2008). For predicting dropout, consistency was indeed highest when participants saw the statement in narrative form. However, for predicting GPA, the differences were negligible, so it seems unlikely that consistency differences fully explain these unexpected findings. Moreover, we do not have an explanation for why consistency differences would have occurred.

As expected, participants reported higher perceived autonomy when they used their own versus others' quantifications of personal statements to make predictions, but no evidence was found for higher confidence. Nevertheless, use intentions were higher for self-produced quantifications. However, the weights assigned to self-produced and other-produced quantifications were identical, and, in line with that finding, no predictive accuracy differences were found. The findings thus lack support for the hypothesis that self-produced quantifications would be weighted more heavily and would subsequently reduce predictive accuracy. These findings do not support the recommendation that those involved in information combination should not be involved in data collection (Kuncel, 2018).

8.1 | Strengths and limitations

We used a prediction task using stimulus material from a real selection context. To ensure a representative design (Brunswik, 1956), we asked participants to make predictions not only based on the stimuli of interest, but also other valid and commonly used predictors, and using random samples of 192 different applicants. Furthermore, two different criteria were used, one continuous and one binary. Using stimulus material and criterion measures from the context of college admissions ensured that the undergraduate participants were familiar with both the predictors and the outcome measures they were asked to predict, as is representative for hiring and admissions decisions in practice. Nevertheless, conducting a similar study with industrial-organizational psychologists, HR-practitioners, or hiring managers, using stimulus material that is representative of the prediction tasks they engage in, would be very valuable. Our sample was also heavily dominated by female participants, even more so than would be expected based on the population from which participants were sampled.

Another limitation is that we only tested our hypotheses using a within-subjects design. A between-subjects design could yield different results, because a within-subjects design allows joint or relative evaluation, while a between-subjects design requires separate evaluation, which is often more difficult (Highhouse et al., 2017; Hsee et al., 1999). However, the comparative or joint mode is likely most representative for situations in which decisions on the adoption of selection practices are made in practice (Hsee & Zhang, 2004), which is a benefit when investigating measures such as use intentions (Nolan et al., 2020).

Moreover, it would be valuable to collect more judgments per decision maker in future research, to allow building stable judgmental bootstrapping models on the individual level, rather than on the group level.

To investigate validity and how participants weighted the predictors, we had to ask participants to provide quantitative ratings of the statements in the narrative condition. Consequently, even though participants processed the information in narrative form when making their predictions, they also knew their own quantitative rating. Producing this quantitative rating could have influenced their predictions and is not fully representative of how narrative information is typically used in practice. This limitation is unavoidable given the data needed to investigate our hypotheses.

We also used just one type of predictor that was varied in terms of format and source. Using other commonly used instruments such as (unstructured) interviews or individual assessment reports, which often contain both quantitative scores and narrative descriptions (e.g., Morris et al., 2015) could yield different results. For example, especially unstructured interviews have been hypothesized to allow “crafting the narrative” toward a coherent story by asking questions that confirm existing theories about the applicant (Dana et al., 2013). Additionally, we did not fully cross information format with information source. Using different predictors would also allow including a condition where the original, raw information is processed by some other than the decision maker and their interpretation is presented in narrative form (e.g., a narrative assessment report written by someone else).

8.2 | Conclusion, implications and future research

It should be noted that, while the hypothesized attenuating effect of narrative or self-produced quantitative information on predictive accuracy was not detected, including narrative or self-produced quantifications also did not seem to have positive effects on predictive accuracy, despite being perceived as more informative and yielding higher use intentions. In addition, for narrative compared to quantified information, the results on weights assigned to the predictors demonstrate the potential for attenuating effects on accuracy due to overweighting less valid information, even when that effect was not significantly present in this study. Overall, the jury is still out on whether, when and how information format affects information utilization and predictive accuracy. Therefore, no specific practical recommendations on what information format or quantification source to use can be provided at this point.

In general, the effect of information format on utilization and predictive accuracy is a relevant topic for future research (Kuncel, 2018). In this study, we focused on the possible detrimental effects of certain formats on predictive accuracy through overweighting predictively nonvalid, but face valid information. However, it would be interesting to investigate if certain presentation formats or collection sources could *enhance* the utilization of predictively valid but less face valid predictors, such as scores on standardized tests (see Zhang et al., 2019). For example, perhaps decision makers are more likely to utilize scores on standardized tests or assessment center ratings when they were involved in their administration, selection, or design.

DATA AVAILABILITY STATEMENT

Part of the data that support the findings of this study are available on request from the corresponding author. Those data (demographics, admission test scores, high school GPA, GPA and drop out) are not publicly available due to privacy or ethical restrictions. The other data is available via the OSF project page.

ORCID

A. Susan M. Niessen http://orcid.org/0000-0001-8249-9295
Edgar E. Kausel https://orcid.org/0000-0002-7181-0954
Marvin Neumann http://orcid.org/0000-0003-0193-8159

ENDNOTES

[1] Defining dilution as the underutilization of valid information in the presence of nonvalid information is somewhat controversial. Originally, the dilution effect was studied as the effect of nonvalid information on the extremity of judgments and predictions (Kemmelmeier, 2004; Nisbett et al., 1981; Troutman & Shanteau, 1977), and the definition of the dilution effect often formally includes this effect on the extremity; it is the *extremity* of judgments and predictions that is diluted (e.g., Highhouse et al., 2021, this issue). Later studies (Dana et al., 2013; Hall et al., 2007; Kausel et al., 2016; Waller & Zimbelman, 2003) also investigated the effect of access to nonvalid information on predictive accuracy, expecting that access to nonvalid information would reduce the utilization of valid information, resulting in reduced predictive accuracy. Some referred to this effect as dilution as well. However, whether the latter should be labeled a dilution effect is a matter of dispute (e.g., Dalal et al., 2020; Highhouse et al., 2021, this issue). We propose that the dilution effect should be defined as a weakening of the impact of valid information on judgments and predictions as a result of the presence of nonvalid information (Nisbett et al., 1981; Waller & Zimbelman, 2003). In turn, this dilution effect can affect the extremity and/or the accuracy of judgments and predictions. Hence, it is *the information used to make judgments and predictions* that is diluted by adding nonvalid information. This definition aligns with explanations presented in the original article by Nisbett et al. (1981, p. 251) “The effect of nondiagnostic information, then, might be to “dilute” the implications of diagnostic information…” and more recently in Highhouse et al. (2021, p. 1, this issue) “The dilution effect occurs when the presence of nondiagnostic information weakens the potency of diagnostic information in predictions about future success”. Notably, Highhouse et al. (2021, this issue), found that the effect of dilution on the *extremity* of predictions did not replicate in a performance prediction context. However, the effects of dilution on accuracy are not expected to be (solely) driven by effects on extremity (Highhouse et al., 2021). In this study, we focus on the effects of dilution on predictive accuracy.

[2]Tabulated descriptive statistics are provided in Supporting Information S2.

REFERENCES

Allen, M., & Preiss, R. W. (1997). Comparing the persuasiveness of narrative and statistical evidence using meta-analysis. *Communication Research Reports*, *14*, 125–131. https://doi.org/10.1080/08824099709388654

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*, 291–304. https://doi.org/10.1111/j.1468-2389.2010.00512.x

Armstrong, J. S. (2001). Judgmental bootstrapping: Inferring experts' rules for forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting. International series in operations research & management science* (pp. 171–192). Springer. https://doi.org/10.1007/978-0-306-47630-3_9

Beach, L. R. (2010). *The psychology of narrative thought*. Xlibris.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Boswell, W. R., Roehling, M. V., LePine, M. A., & Moynihan, L. M. (2003). Individual job-choice decisions and the impact of job attributes and recruitment practices: A longitudinal field study. *Human Resource Management*, *42*, 23–37. https://doi.org/10.1002/hrm.10062

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Dalal, D. K., Sassaman, L., & Zhu, X. (2020). The impact of nondiagnostic information on selection decision making: A cautionary note and mitigation strategies. *Personnel Assessment and Decisions*, *6*, 54–64. https://doi.org/10.25035/pad.2020.02.007

Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes*, *112*, 11–23. https://doi.org/10.1016/j.obhdp.2009.11.007

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

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Deci, E. L., & Ryan, R. M. (1991). A motivational approach to self: Integration in personality. In R. A. Dienstbier (Ed.), *Nebraska symposium on motivation, 1990: Perspectives on motivation* (pp. 237–288). University of Nebraska Press.

Dipboye, R. L. (1997). Structured selection interviews: Why do they work? Why are they underutilized? In N. Anderson, & P. Herriott (Eds.), *International handbook of selection and assessment* (pp. 455–474). Wiley.

Dipboye, R. L., & Jackson, S. L. (1999). Interviewer experience and expertise effects. In W. Eder, & M. M. Harris (Eds.), *The employment interview handbook* (pp. 259–278). Sage Publications Inc.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, *73*, 422–432. https://doi.org/10.1037/h0029230

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*, 1–27. https://doi.org/10.18637/jss.v017.i01

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, *103*, 277–290. https://doi.org/10.1016/j.obhdp.2007.01.003

Highhouse, S. (1997). Understanding and improving job-finalist choice: The relevance of behavioral decision research. *Human Resource Management Review*, *7*, 449–470. https://doi.org/10.1016/S1053-4822(97)90029-2

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

Highhouse, S., Brooks, M. E., Nesnidol, S., & Sim, S. (2017). Is a .51 validity coefficient good? Value sensitivity for interview validity. *International Journal of Selection and Assessment*, *25*, 383–389. https://doi.org/10.1111/ijsa.12192

Highhouse, S., Freier, L. M., Stevenor, B. A., Shea, M. A., Childers, M., & Melick, S. R. (2021). Failure to replicate the basic dilution effect in performance prediction. *International Journal of Selection and Assessment*. Published online September 20, 2021. https://doi.org/10.1111/ijsa.12344

Highhouse, S., & Kostek, J. A. (2013). Holistic assessment for selection and placement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology. Test theory and testing and assessment in industrial and organizational psychology* (Vol. 1, pp. 565–577). *American Psychological Association*. https://doi.org/10.1037/14047-031

Hsee, C. K., Loewenstein, G., Blount, S., & Bazerman, M. (1999). Preference reversals between joint and separate evaluations of options: A theoretical analysis. *Psychological Bulletin*, *125*(5), 576–590. https://doi.org/10.1037/0033-2909.125.5.576

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology*, *86*, 680–695. https://doi.org/10.1037/0022-3514.86.5.680

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology. *International Journal of Selection and Assessment*, *22*, 297–309. https://doi.org/10.1111/ijsa.12078

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*, 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Kausel, E. E., Culbertson, S. S., & Madrid, H. P. (2016). Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring decisions. *Organizational Behavior and Human Decision Processes*, *137*, 27–44. https://doi.org/10.1016/j.obhdp.2016.07.005

Kemmelmeier, M. (2004). Separating the wheat from the chaff: Does discriminating between diagnostic and nondiagnostic information eliminate the dilution effect? *Journal of Behavioral Decision Making*, *17*, 231–243. https://doi.org/10.1002/bdm.473

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 343–346. https://doi.org/10.1111/j.1754-9434.2008.00059.x

Kuncel, N. R. (2018). Judgment and decision making in staffing research and practice. In D. S. Ones, N. Anderson, C. Viswesvaran, & H. K. Sinangil (Eds.), *The SAGE handbook of industrial, work & organizational psychology: Personnel psychology and employee performance* (Vol. 1, 2nd ed., pp. 474–488). Sage Reference.

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*, 1060–1072. https://doi.org/10.1037/a0034156

Lievens, F., & De Paepe, A. (2004). An empirical investigation of interviewer-related factors that discourage the use of high structure interviews. *Journal of Organizational Behavior*, *25*, 29–46. https://doi.org/10.1002/job.246

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*, 629–650. https://doi.org/10.1093/jcr/ucz013

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*, 6–27. https://doi.org/10.1037/met0000086

Morris, S. B., Daisley, R. L., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology*, *100*, 5–20. https://doi.org/10.1037/a0036938

Neumann, M., Hengeveld, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2021). Education increases decision-rule use: An investigation of education and incentives to improve decision making. *Journal of Experimental Psychology: Applied*, 1–13. https://doi.org/10.1037/xap0000372

Neumann, M., Niessen, A. S. M., & Meijer, R. R. (2021). Implementing evidence-based assessment and selection in organizations: A review and an agenda for future research. *Organizational Psychology Review*, *11*, 205–239. https://doi.org/10.1177/2041386620983419

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Applying organizational justice theory to admission into higher education: Admission from a student perspective. *International Journal of Selection and Assessment*, *25*, 72–84. https://doi.org/10.1111/ijsa.12161

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLOS One*, *13*(6), e0198746. https://doi.org/10.1371/journal.pone.0198746

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*, 248–277. https://doi.org/10.1016/0010-0285(81)90010-4

Nolan, K. P., Dalal, D. K., & Carter, N. (2020). Threat of technological unemployment, use intentions, and the promotion of structured interviews in personnel selection. *Personnel Assessment and Decisions*, *6*, 38–53. https://doi.org/10.25035/pad.2020.02.006

Nolan, K. P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance*, *27*, 328–346. https://doi.org/10.1080/08959285.2014.929691

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, *62*, 189–206. https://doi.org/10.1037/0022-3514.62.2.189

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115. https://doi.org/10.1037/a0022658

Revelle, W. (2019). *Psych: Procedures for psychological, psychometric, and personality research*. R package version 1.9.12. https://CRAN.R-project.org/package=psych

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology*, *40*, 455–488. https://doi.org/10.1111/j.1744-6570.1987.tb00610.x

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management*, *41*, 149–174. https://doi.org/10.1002/hrm.10029

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*, 1435–1447. https://doi.org/10.1037/apl0000236

Sanfey, A., & Hastie, R. (1998). Does evidence presentation format affect judgment? An experimental evaluation of displays of data for judgments. *Psychological Science*, *8*, 99–103. https://doi.org/10.1111/1467-9280.00018

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*, 262–274. https://doi.org/10.1037/0033-2909.124.2.262

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* (pp. 57–79). Routledge/Taylor & Francis Group.

Steiger, J. H. (1980). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research*, *15*, 335–352. https://doi.org/10.1207/s15327906mbr1503_7

Tonidandel, S., & LeBreton, J. M. (2015). RWA Web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology*, *30*, 207–216. https://doi.org/10.1007/s10869-014-9351-z

Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior & Human Performance*, *19*, 43–55. https://doi.org/10.1016/0030-5073(77)90053-8

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology*, *61*, 871–925. https://doi.org/10.1111/j.1744-6570.2008.00133.x

Waller, W. S., & Zimbelman, M. F. (2003). A cognitive footprint in archival data: Generalizing the dilution effect from laboratory to field settings. *Organizational Behavior and Human Decision Processes*, *91*, 254–268. https://doi.org/10.1016/S0749-5978(03)00024-4

Winterbottom, A., Bekker, H. L., Conner, M., & Mooney, A. (2008). Does narrative information bias individual's decision making? A systematic review. *Social Science & Medicine*, *67*, 2079–2088. https://doi.org/10.1016/j.socscimed.2008.09.037

Witt, J. K. (2019). Graph construction: An empirical investigation on setting the range of the y-axis. *Meta-Psychology*, *3*, 1–20. https://doi.org/10.5626/MP.2018.895

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*, 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260–281. https://doi.org/10.1006/obhd.2000.2909

Yu, M. C. (2018). *Viewing expert judgment in individual assessments through the Lens Model: Testing the limits of expert information processing* (Doctoral dissertation). Minneapolis, MN: University of Minnesota.

Zhang, D. C., Zhu, X., Ritter, K. J., & Thiele, A. (2019). Telling stories to communicate the value of the pre-employment structured job interview. *International Journal of Selection and Assessment*, *27*, 299–314. https://doi.org/10.1111/ijsa.12264

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Niessen, A. S. M., Kausel, E. E., & Neumann, M. (2022). Using narratives and numbers in performance prediction: Attitudes, confidence and validity. *International Journal of Selection and Assessment*, 30, 216–229. https://doi.org/10.1111/ijsa.12364