The autonomy-validity dilemma in mechanical prediction procedures

Neumann, Marvin; Niessen, A. Susan M.; Tendeiro, Jorge N.; Meijer, Rob R.

Published in:
Journal of Behavioral Decision Making

DOI:
10.1002/bdm.2270

Publication date:
2022

Received: 2 March 2021 | Revised: 12 November 2021 | Accepted: 13 November 2021

DOI: 10.1002/bdm.2270

RESEARCH ARTICLE

WILEY

The autonomy-validity dilemma in mechanical prediction procedures: The quest for a compromise

Marvin Neumann[1] | A. Susan M. Niessen[1] | Jorge N. Tendeiro[2] | Rob R. Meijer[1]

[1]Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands

[2]Office of Research and Academia-Government-Community Collaboration, Education, Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Hiroshima, Japan

Correspondence
Marvin Neumann, Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.
Email: m.neumann@rug.nl

Abstract

A robust finding in psychological research is that combining information with a mechanical rule results in more valid predictions than combining information holistically in the mind. Nevertheless, information is typically combined holistically in practice, resulting in suboptimal predictions and decisions. Earlier research showed that decision makers are more likely to use mechanical prediction procedures when they retain autonomy in the decision-making process. However, it remains largely unknown how different autonomy-enhancing features affect predictive validity. Therefore, in two preregistered studies (total $N = 342$), we investigated if and how prediction procedures can be designed such that they satisfy decision makers' autonomy needs and acceptance without reducing predictive validity. Based on archival application data from a university admission procedure, participants predicted applicants' first-year grade point average and chance of dropout. The results of Bayesian analyses showed that participants preferred prediction procedures in which they retained autonomy by choosing consistent predictor weights of a mechanical rule or by holistically adjusting the predictions of an optimal regression model. In general, these prediction procedures resulted in slightly higher predictive validity compared with fully holistic prediction. Providing participants with predictor validity information slightly increased predictive validity when participants could choose predictor weights but not when making holistic predictions or adjusting optimal model predictions. Giving decision makers a role in designing mechanical rules through choosing weights based on explicit predictive validity information could help promote the implementation and validity of mechanical prediction in practice.

KEYWORDS

algorithm aversion, autonomy-validity dilemma, decision aid, decision-making, holistic prediction, mechanical prediction, selection

1 | INTRODUCTION

One of the most robust findings in psychological research is that mechanical prediction results in higher predictive validity than holistic prediction (Ægisdóttir et al., 2006; Dana & Thomas, 2006; Ganzach et al., 2000; Grove et al., 2000; Kuncel et al., 2013; Meehl, 1954a; Yu & Kuncel, 2020). In holistic (or clinical) prediction, decision makers combine information "in their mind," often intuitively. In mechanical (or actuarial, statistical) prediction, decision makers combine information with a consistent predefined mechanical rule or formula (Grove et al., 2000; Meehl, 1954a). Imagine admission officers selecting students from a larger pool of applicants based on a standardized test score, high school grade point average (GPA), and a personal statement rating. They could make performance predictions by integrating this information in their mind (holistic prediction) or by applying a consistent rule in which each piece of information is given an explicit weight (mechanical prediction). Those weights can be equal or different and can be based on analyses of primary data, meta-analyses, subject matter experts, or decision makers themselves (Bobko et al., 2007; Murphy et al., 2013).

Holistic prediction is often less valid than mechanical prediction for two main reasons; *inconsistently* and *inaccurately* weighting information (Karelaia & Hogarth, 2008). Inconsistency seems to be the main driver for lower predictive validity of holistic prediction, as even randomly chosen but consistent weights (with the correct sign) outperform holistic prediction (Dawes & Corrigan, 1974; Yu & Kuncel, 2020). However, in a study of hiring managers predicting applicants' job performance, Kausel et al. (2016) found that inaccurate weighting of predictors rather than inconsistency was the main driver of reduced prediction accuracy, compared with using an optimal mechanical rule.

Although studies demonstrating the superiority of mechanical prediction have been published for decades (Grove et al., 2000; Meehl, 1954a; Sarbin, 1943; Sawyer, 1966), there still exists a large gap between science and practice. Even in highly consequential contexts such as admission to higher education, personnel selection, and clinical treatment selection, decision makers underutilize mechanical prediction (Arkes, 2008; Highhouse, 2008; Kirch, 2012; Neumann, Niessen, et al., 2021; Ryan et al., 2015; Ryan & Sackett, 1987; Vrieze & Grove, 2009). This phenomenon has also been referred to as algorithm aversion (Dietvorst et al., 2015).

A major reason why decision makers underutilize mechanical rules is that it restricts their autonomy (Neumann, Niessen, et al., 2021; Nolan & Highhouse, 2014). Some studies found that providing decision makers with autonomy increased their (intended) use of mechanical rules (Dietvorst et al., 2018; Nolan & Highhouse, 2014). However, enhancing decision makers' autonomy to increase the use of mechanical prediction should not significantly attenuate predictive validity. Therefore, in two preregistered studies (https://osf.io/86jfb/registrations) about predicting human performance, we investigated if and how prediction procedures can be designed such that they satisfy autonomy needs and are accepted by decision makers without losing predictive validity.

2 | WHY DECISION MAKERS PREFER HOLISTIC PREDICTION

Decision makers prefer holistic over mechanical prediction (Eastwood et al., 2012) for a number of reasons, such as unawareness of research findings on evidence-based decision-making (Neumann, Hengeveld, et al., 2021; Rynes, 2012), stakeholders' appreciation for holistic prediction (Nolan et al., 2016), and a preference for riskier prediction methods in uncertain decision-making contexts (Dietvorst & Bharti, 2020). Furthermore, decision makers are overly confident in their holistic predictions (Kausel et al., 2016; Sieck & Arkes, 2005), have the desire to take an individual's uniqueness and context into account (Longoni et al., 2019; Newman et al., 2020), and wrongly believe that they can account for valid exceptions to a mechanical rule (Dietvorst et al., 2018; Guay & Parent, 2018; Hoffman et al., 2017; Meehl, 1954b). Overall, a main reason seems to be that holistic prediction better satisfies decision makers' fundamental needs (Neumann, Niessen, et al., 2021). According to self-determination theory, people have three fundamental needs: autonomy, competence, and relatedness (Deci & Ryan, 2000). Autonomy needs are satisfied when people experience choice and control over processes. Competence and relatedness refer to people's need to experience efficacy and a connection with and care for others, respectively (Deci et al., 2001). All three needs may play a role in the implementation of mechanical prediction (Nolan, 2013). Yet, among existing studies that have focused on interventions to implement mechanical prediction procedures, the most promising interventions increased decision makers' autonomy in mechanical prediction in some way (Kaplan et al., 2001; Neumann, Niessen, et al., 2021). Given these prior findings and its solid theoretical basis, we focused on autonomy as the main variable of interest in this paper.

3 | SATISFYING AUTONOMY IN MECHANICAL PREDICTION

3.1 | Choosing predictor weights

In line with self-determination theory, Nolan and Highhouse (2014) found that participants experienced more autonomy in holistic than in mechanical prediction when responding to hypothetical hiring scenarios. Furthermore, participants reported higher intentions to use mechanical prediction if they could choose predictor weights (autonomy), compared with if they had to use organizationally prescribed predictor weights (no autonomy). So, giving decision makers control over the design of the mechanical rule by choosing the weights increased use intentions of mechanical rules through higher perceived autonomy.

Decision makers could choose predictor weights in two ways. They may choose one general set of predictor weights that will be applied to all applicants to be judged (as participants imagined in the second study by Nolan & Highhouse, 2014). This increases consistency and hence results in more valid predictions than holistic prediction (Yu & Kuncel, 2020), and decision makers' autonomy is retained through

control over the predictor weights (Nolan & Highhouse, 2014). Alternatively, decision makers could choose different predictor weights for each applicant to be judged. This allows decision makers to account for suspected exceptions to the linear rule, such as interactions. For example, decision makers may weight an applicant's test score less heavily when their high school grade was excellent (for a similar example, see MacDonald, 2020). However, this approach still requires decision makers to choose explicit weights, which could increase consistency compared with holistic prediction if decision makers would be consistent in a majority of cases (where they do not expect rule exceptions), while holistic predictions for these applicants may still be inconsistent due to unintentional factors such as fatigue and inattention. This approach contains components of mechanical and holistic prediction; the combination of information does not happen "in the mind," as in holistic prediction but is based on explicit weights. However, because weights are not consistent across individuals, this approach does not meet the definition of mechanical prediction.

3.2 | Adjusting the result of a mechanical rule

Autonomy is also retained when decision makers can holistically adjust the result after applying a mechanical rule. This procedure, which is also called clinical synthesis (Kuncel, 2018; Sawyer, 1966), allows decision makers to account for suspected rule exceptions more directly. Although not explicitly based on self-determination theory, Dietvorst et al. (2018) conducted a series of experiments in which participants' autonomy in using a mechanical rule to make standardized test performance predictions was varied. The results showed that participants who could adjust the rule's predictions were more confident in and more satisfied with their predictions, chose to use the rule more often, and made more accurate predictions as a result, compared with participants who could not adjust the rule's predictions.

Although predictive validity decreases when decision makers holistically adjust mechanical rule predictions (Guay & Parent, 2018; Hanson & Morton-Bourgon, 2009; Hoffman et al., 2017; Neumann, Hengeveld, et al., 2021), there is some evidence that adjusting rule predictions still results in higher predictive validity than pure holistic prediction (Dietvorst et al., 2018). This is because a rule prediction based on appropriate weights provides a valid "anchor" (Dietvorst et al., 2018; Tversky & Kahneman, 1974).

In sum, the results of Nolan and Highhouse (2014) and Dietvorst et al. (2018) suggest that the use of mechanical rules can be increased by granting decision makers autonomy over predictor weights or over the rule prediction. However, it remains unclear which of these procedures yields the most favorable results in terms of user perceptions, intentions, and predictive validity.

3.3 | Aim and contribution

The aim of the current studies was to investigate the effects of different kinds of autonomy-enhancing features in using a mechanical rule on (1) user perceptions and intentions and (2) predictive validity, to see if a satisfactory compromise between optimizing autonomy and validity can be established. The present studies extend the studies by Nolan and Highhouse (2014) and Dietvorst et al. (2018) in three ways. First, Nolan and Highhouse (2014) investigated whether autonomy in determining predictor weights affected use intentions for mechanical rules based on vignettes describing different procedures. We instead asked our participants to make actual predictions based on real data. Second, we also investigated the effect of autonomy enhancement on predictive validity: Enhancing decision makers' autonomy should not result in a large reduction in predictive validity to be of any practical value. Third, to our knowledge, no studies have jointly investigated autonomy in designing the rule and autonomy through adjusting rule predictions. In the current studies, we included both prediction procedures, to investigate whether attitudes and validity depend on the stage in the decision-making process in which decision makers have autonomy (Burton et al., 2020). In sum, we respond to the calls for investigations of how decision makers' acceptance of mechanical prediction procedures can be increased without substantially reducing predictive validity (Burton et al., 2020; Kuncel, 2018; Kuncel et al., 2013; Neumann, Niessen, et al., 2021; Ryan et al., 2015; Sackett & Lievens, 2008).

4 | STUDY 1

We employed a within-subjects design to ensure sufficient statistical power and to allow for a joint evaluation of the different prediction procedures. This approach yields the most representative measure of use intentions, because in practice, decisions to use certain prediction procedures are made by comparing multiple procedures jointly rather than separately (Highhouse et al., 2017; Hsee & Zhang, 2004; Nolan et al., 2020). Furthermore, given that many decision makers in practice are unaware of the distinction between holistic and mechanical prediction, a within-subjects design should improve the evaluability of an attribute that is difficult to evaluate due to a lack of knowledge (Hsee & Zhang, 2010).

In the first four conditions, each participant predicted the first-year GPA and the chance of dropout of five applicants to a psychology undergraduate program, based on three predictors: high school GPA, an admission test score, and a personal statement. In the first condition, participants made holistic predictions. This condition was included so the other conditions could be compared with the prediction procedure typically used in practice. In the second and third conditions, participants could design a mechanical rule by choosing general predictor weights (general weights condition) or by choosing applicant-specific predictor weights (individual weights condition), respectively. Including these conditions allowed us to test whether participants would prefer designing rules tailored to the unique individual (individual weights) over designing one consistent rule (general weights). This preference was expected because people often think information should be interpreted in the context of all other information (Grove & Meehl, 1996) and because decontextualization seems

to be a driver of algorithm aversion (Newman et al., 2020). In the fourth condition—the adjustment condition—participants could adjust the prediction of a statistical model (see Dietvorst et al., 2018), in a similar way as in typical advice-taking studies (Önkal et al., 2009). Including this condition allowed us to investigate whether perceptions and use intentions and predictive validity differ depending on whether participants design the prediction process (choosing general weights) or have the final say at the end (adjusting optimal model predictions), which are two qualitatively different types of autonomy-enhancing features. Lastly, in the fifth condition, participants did not make predictions themselves (no autonomy) but had to rely on optimal regression model predictions which they could not adjust (optimal model condition). This condition was included to compare the previous four conditions with this optimal but least utilized procedure in practice. Based on the findings of Nolan and Highhouse (2014) and Dietvorst et al. (2018), we formulated the following expectations:

> **Hypothesis 1.** Perceived autonomy will decrease from the holistic condition (most autonomy), individual weights condition, and general weights condition to the optimal model condition (least autonomy).

We had no specific hypothesis on how the adjustment condition would compare to the two weighting conditions, because participants could adjust optimal model predictions and hence had complete autonomy. Yet, seeing the model prediction may reduce autonomy perceptions compared with making holistic predictions. Therefore, we only expected that perceived autonomy will be higher compared with the optimal model condition and lower compared with the holistic condition. For brevity, Hypothesis 1 is only formulated for perceived autonomy. We expected the same results for use intentions (H2), confidence (H3), and satisfaction (H4).

> **Hypothesis 5.** Predictive validity will increase from the holistic condition (lowest predictive validity), individual weights condition, and general weights condition to the optimal model condition (highest predictive validity).

We had no specific hypothesis on how the adjustment condition would compare to the two weighting conditions but expected lower predictive validity than in the optimal model condition and higher predictive validity than in the holistic condition.

5 | METHOD

5.1 | Participants

A power analysis[1] for the omnibus *F*-test of a repeated measures Analysis of Variance (ANOVA, Hypotheses 1–4) resulted in a required sample size of $N = 40$ (assuming $1\text{-}\beta > .8$, $\alpha = .05$, number of measurements = 5, nonsphericity correction $\varepsilon = 1$, and a moderate effect size, $f = 0.25$, consistent with previous research, Nolan & Highhouse, 2014). Furthermore, since we had no a priori estimate of the correlation among repeated measures, we set this correlation equal to zero, which results in the most conservative sample size estimation). To be conservative and counteract inflated effect sizes in the published literature (Schäfer & Schwarz, 2019), we aimed to collect minimally 50 participants and maximally as many as signed up before the planned closing date of data collection (see the preregistration protocol, https://osf.io/fjbvw).

Many more participants than minimally required signed up before the planned closing date of data collection, which resulted in a total of $N = 153$ first-year psychology students who consented to take part in this study for course credit. Three participants failed one or both of two attention checks (see Appendix A) and were excluded. The final dataset consisted of $N = 150$[2] participants (17% male, 83% female, <1% did not want to disclose) who ranged in age from 18 to 39 years ($M = 19.85$, $SD = 2.52$). Most participants had the Dutch nationality (97%). The other four participants had German, Brazilian, Russian, and Syrian nationalities. This study was approved by the ethics committee of the Heymans institute for psychological research (code: 18255 – S).

5.2 | Materials

Participants were presented with anonymized archival data from applicants to the psychology undergraduate program of the University of Groningen. Application data of 192 Dutch applicants were used as stimulus material (Niessen et al., 2018). From this pool, five applicants were randomly allocated (without replacement within participants) to the four conditions in which participants made predictions (i.e., all conditions except the optimal model condition). Applicants were evenly sampled so that every applicant was judged. So, applicants were assigned to participants until all 192 applicants were sampled, after which a new sampling round started. Participants were presented with each applicant's high school GPA, an admission test score, and a personal statement. We used these predictors because they are commonly used in college admission procedures (Davis et al., 2018; Niessen et al., 2018).

5.3 | High school GPA

High school GPA was the mean of all final grades obtained at the end of secondary education (vwo) and was measured on the Dutch 10-point grading scale (1 is lowest, 10 is highest). High school GPA ranged from 5.90 to 8.55 ($M = 6.61$, $SD = 0.46$) and was a good predictor of first-year GPA ($r = .45$) and a moderate predictor of dropout ($r = -.23$) in this sample.

5.4 | Admission test

Applicants had to study two chapters from an introductory psychology book and then took an admission test of 40 multiple choice items

that covered the study material. "Number of items correct" was transformed to the Dutch 10-point grading scale and ranged from 2.2 to 9.7 ($M = 6.42$, $SD = 1.77$). In this sample, the admission test was a good predictor of first-year GPA ($r = .40$) and a moderate predictor of dropout ($r = -.27$).

5.5 | Personal statement

Before making predictions, participants also rated on a 7-point scale (1 = *very unmotivated*, 7 = *very motivated*) applicants' personal statements, in which applicants expressed their motivation to study psychology at the university in a maximum of 250 words. The mean rating was $M = 5.47$ ($SD = 1.08$). For the calculation of predicted scores within conditions, personal statement ratings were transformed to the Dutch 10-point grading scale. Research suggests that inter-rater reliability of personal statements is low (GlenMaye & Oakes, 2002; Murphy et al., 2009). Indeed, inter-rater reliability was also low in this study ($ICC = .24$). The personal statement rating was a poor predictor of first-year GPA across conditions ($\bar{r} = .05$) and of dropout ($\bar{r} = .02$).

5.6 | Procedure and conditions

Participants completed the study in a lab space at the university in groups of up to 16 participants simultaneously. The experimenter informed participants about the opportunity to earn a reward (up to €5), depending on their prediction accuracy. The reward served to simulate accountability and to encourage effort, as is common in this area of research (e.g., Dietvorst et al., 2018). All other instructions were presented on a computer screen. The participants were informed that they would first rate personal statements and then predict applicants' first-year GPA and chance of dropout in the first year of the psychology program. Participants received predictor validity information. They were informed that high school GPA and the admission test were "good predictors," while the personal statement was a "poor predictor" of the outcomes. Participants also received summary statistics (mean, minimum, and maximum) for the predictors high school GPA and the admission test score, and the criteria first-year GPA and dropout as reference.

In the *holistic* condition, participants saw the applicants' data (i.e., high school GPA, admission test score, personal statement in qualitative form, and their quantitative rating) and holistically predicted applicants' first-year GPA and the chance of dropout.

In the *individual weights* condition, participants saw the applicants' data and assigned percentage weights to the predictors *for each of the five applicants individually*. So, participants could assign different predictor weights to different applicants. To calculate the resulting predictive score, each predictor score was multiplied by the individual weight that a participant assigned to this predictor.

In the *general weights* condition, participants assigned percentage weights to the predictors once, which applied to *all of the five applicants* to be judged in this condition. Each predictor score was multiplied by the weight that a participant assigned to this predictor.

Additionally, we also wanted to explore whether thinking about predictor importance (by choosing a set of general weights) would by itself result in more consistent and hence more valid holistic predictions that immediately follow after choosing weights, compared with predictions in the holistic condition. Therefore, after participants chose general weights and responded to the attitudinal measures, they again judged five applicants holistically, in the same manner as in the holistic condition. We present the additional details (Section S1) and the results (Section S2) in the supporting information.

In the *adjustment* condition, participants saw the applicants' data and a regression model prediction of applicants' first-year GPA and chance of drop out. Participants were told that the model was based on an applicant's high school GPA and admission test score.[3] Furthermore, they were informed that the model was quite accurate in predicting the criteria and that it most likely gave the best prediction, better than their own, holistic prediction. Participants could holistically adjust the model prediction. If they did not want to adjust it, they simply reproduced the model prediction.

In the *optimal model* condition, participants did not make predictions themselves but imagined that they had to use the regression model predictions based on the same model as in the adjustment condition, without the possibility to adjust it in any way.

Each participant completed each condition. The order of the conditions was counterbalanced,[4] except for the optimal model condition. This condition was always presented last because the adjustment condition always had to be presented before the optimal model condition. The instructions to participants per condition are shown in Appendix B. At the end of each condition, participants answered questions regarding their intentions to use, their confidence in, their satisfaction with, and their perceived autonomy in the prediction approach.

Finally, participants imagined that their task was to select applicants that would achieve a high first-year GPA and would be unlikely to drop out for the upcoming academic year. Participants ranked the different prediction procedures according to which procedure they were most likely to use (rank 1 = *most likely*, rank 5 = *least likely*). Participants were informed that their reward they had earned up to this point would be increased by 20% if they chose the procedure that yielded the most accurate predictions (we assumed this to be the optimal model condition, as indeed turned out to be the case). The descriptions of the five procedures were displayed in random order. The median time it took to complete the study was 39 min.

5.7 | Attitudinal measures

Perceived autonomy was measured with a translated version of the six-item scale ($\alpha = .91$) from Nolan and Highhouse (2014). We adapted the three-item scale from Nolan and Highhouse (2014) to assess the extent to which participants intend to use a certain approach for future admission decisions ($\alpha = .83$). Confidence and satisfaction were assessed with one-item measures based on

Dietvorst et al. (2018). Participants responded to these measures on 5-point Likert scales. More details about the measures can be found online (https://osf.io/86jfb/).

5.8 | Measures used for predictive validity hypotheses

5.8.1 | First-year GPA

Participants indicated the predicted first-year GPA on a continuous scale (1 = *lowest grade*; 10 = *highest grade*), up to one decimal. The predictive validity was operationalized as the correlation between the predicted and the observed first-year GPA.

5.8.2 | Chance of dropout

Participants indicated the predicted chance that an applicant dropped out of the program on a continuous scale from 0% to 50%. This scale was chosen because the regression model did not return a chance of drop out above 50% for any of the applicants.

5.8.3 | Analytical approach

Dichotomous decision-making based on *p*-values has received much criticism (Cohen, 1994; Cumming, 2014; Wasserstein et al., 2019). Bayesian approaches have been recommended as a suitable alternative because they have several advantages over null hypothesis significance testing (NHST; see Cohen, 1994; Kruschke, 2015; Kruschke et al., 2012; Wagenmakers, 2007). For example, multiple comparisons are not problematic in Bayesian approaches because prior knowledge is incorporated into the model (Berry & Hochberg, 1999; Gelman, 2016; Gelman et al., 2012; Kruschke, 2015). For these reasons, we used a Bayesian approach to analyze our data. We only conducted frequentist tests to analyze the ranks that participants assigned to the prediction procedures because, to our knowledge, there are no Bayesian alternatives for these tests. Moreover, we used Bayesian parameter estimation instead of hypothesis testing (Kruschke & Liddell, 2018) because we were primarily interested in the magnitude and uncertainty of the effects. Furthermore, the null hypothesis is unlikely to be exactly true in most contexts, so testing it against an alternative is much less informative than estimating effect sizes (Kruschke & Liddell, 2018).

The Bayesian approach to parameter estimation allows estimating a distribution that displays the credibility of each possible parameter value (Kruschke, 2011, 2015). This distribution is commonly summarized by a 95% highest density interval (HDI), which is constructed in such a way that it includes the true parameter with probability .95, conditional on the model and observed data. The advantage of HDIs over the commonly misinterpreted frequentist confidence intervals (Morey et al., 2016) is that they can be used to make the intuitive statement that there is a 95% probability that the parameter falls within the boundaries of the HDI, conditional on the model and data at hand. Therefore, we reported parameter estimates and the corresponding 95% HDIs.

6 | RESULTS

All analyses were run in R (4.1.1). We had no permission to share the archival application data due to privacy reasons. Therefore, our publicly available (https://osf.io/86jfb/) R scripts and data allow rerunning all reported analyses except for the results on predictive validity.

6.1 | Attitudinal measures

We ran two-factor mixed-effects Bayesian ANOVAs with the condition as a within-subject factor (fixed effect) and subjects as a random effect, for each of the four dependent measures. The mean differences and the HDIs for the contrasts as stated in Hypotheses 1–4 are reported. Figure 1 displays the observed means per attitudinal measure for each condition. Furthermore, Table 1 shows the means and standard deviations for each attitudinal measure in each condition, and Table 2 shows correlations between the attitudinal measures. As Tables 1 and 2 show, mean patterns and correlations for the confidence and satisfaction measure were very similar to the mean patterns and correlations of the use intentions measure. Therefore, we report the results for confidence and satisfaction and more details of the Bayesian analysis in Section S2 of the supporting information.

6.2 | Perceived autonomy

Figure 1 suggests that perceived autonomy was very similar across prediction procedures, except for using optimal model predictions, where perceived autonomy was much lower compared with all other prediction procedures. As expected, autonomy was highest for making holistic predictions and lowest for using optimal model predictions. Participants experienced less autonomy when choosing individual weights than when making holistic predictions, but this difference was negligible in size ($M_{holistic} - M_{individual} = 0.03$, 95% HDI [−0.10, 0.16], $d = 0.05$). Similarly, the difference in perceived autonomy between choosing general and individual weights was negligible ($M_{individual} - M_{general} = 0.05$, 95% HDI [−0.09, 0.18], $d = 0.07$). However, participants experienced much less autonomy when using optimal model predictions compared with choosing general weights ($M_{general} - M_{optimal} = 0.96$, 95% HDI [0.84, 1.09], $d = 1.18$). Furthermore, perceived autonomy in the adjustment condition fell between the holistic and the optimal model condition. So, Hypothesis 1 was mostly unsupported. Although we observed the hypothesized ordering in perceived autonomy, differences between conditions were negligible, except for the optimal model condition, in which participants

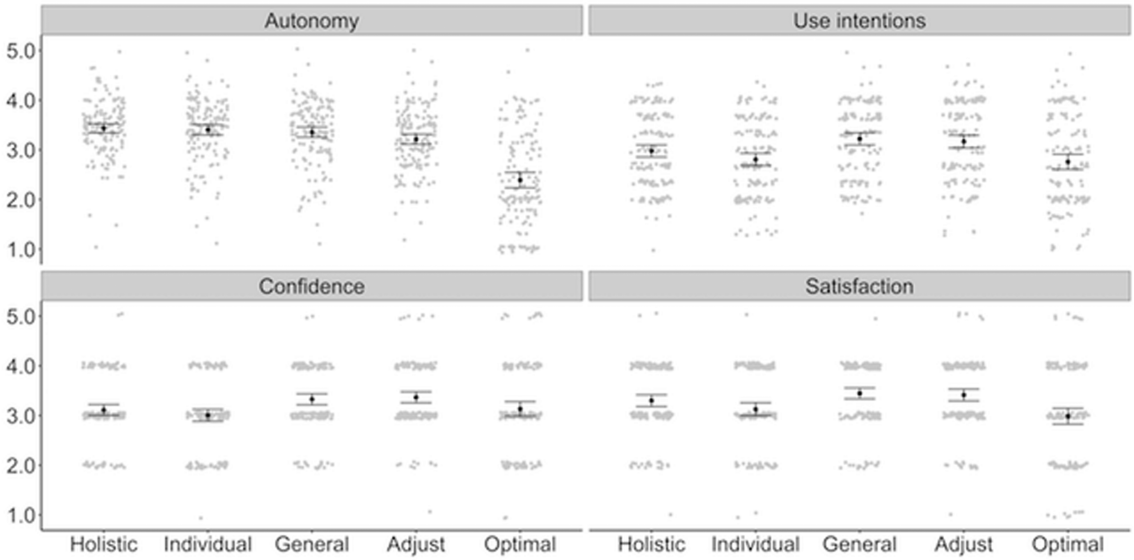

FIGURE 1 Observed means (descriptives) for all conditions per attitude measure in Study 1. *Note*: Error bars represent 95% confidence intervals. Some jittering was added to the plot to improve readability

TABLE 1 Means and standard deviations per condition and attitudinal measure in Study 1

	Autonomy		Use intentions		Confidence		Satisfaction	
Condition	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Holistic	3.43	0.57	2.98	0.76	3.11	0.67	3.30	0.73
Individual	3.40	0.63	2.80	0.78	3.01	0.75	3.13	0.81
General	3.36	0.65	3.20	0.76	3.33	0.69	3.45	0.68
Adjust	3.21	0.64	3.17	0.80	3.37	0.69	3.41	0.73
Optimal	2.39	0.95	2.76	0.94	3.13	0.90	2.99	1.00

TABLE 2 Correlations between attitudinal measures autonomy, use intentions, confidence, and satisfaction in Study 1

Measure	Autonomy	Use intentions	Confidence	Satisfaction
Autonomy	-			
Use intentions	.33*	-		
Confidence	.27*	.63*	-	
Satisfaction	.40*	.70*	.66*	-

Note: Correlations were averaged across conditions using Fisher's z transformation. Correlations between attitudinal measures per condition are reported in the supporting information in Table S1.
*$p < .05$.

experienced much less autonomy compared with all other conditions (Cohen's *d*s > 1.00).

6.3 | Use intentions

We found a number of unexpected results. Use intentions were not highest when making holistic predictions (see Figure 1). Use intentions were noticeably *higher* when choosing general rather than individual weights ($M_{individual} - M_{general} = -0.42$, 95% HDI [−0.58, −0.25], $d = -0.54$) and also slightly higher for adjusting optimal model predictions than for making holistic predictions ($M_{holistic} - M_{adjust} = -0.19$, 95% HDI [−0.35, −0.03], $d = -0.24$). Although, as expected, using optimal model predictions resulted in the lowest use intentions, the difference was practically indistinguishable ($d = 0.05$) from the intentions to use individual weights.

As expected, use intentions were slightly lower for using optimal model predictions than for holistic predictions ($M_{holistic} - M_{optimal} = 0.22$, 95% HDI [0.05, 0.39], $d = 0.26$) and slightly lower for using individual weights than holistic predictions ($M_{holistic} - M_{individual} = 0.17$, 95% HDI [0.01, 0.34], $d = 0.23$). Also, use intentions were markedly lower when using optimal model predictions than when choosing general weights ($M_{general} - M_{optimal} = 0.46$, 95% HDI [0.30, 0.63], $d = 0.54$). In sum, Hypothesis 2 was also mostly

unsupported, as we did not observe the hypothesized decrease in use intentions from the holistic to the optimal condition. However, our results suggest that decision makers welcome prediction procedures in which they can choose general weights or adjust optimal model predictions.

6.4 | Procedure ranking

Participants also ranked the prediction procedures according to which procedure they were most likely to use. The resulting mean preference ranks and rank frequencies per procedure are displayed in Table 3. The majority of participants (56%) preferred using the adjustment procedure for future admission decisions the most. Of interest was that only 10% preferred making holistic predictions the most and few participants preferred choosing predictor weights the most (individual weights 10%; general weights 11%). Also, few participants (13%) preferred optimal model predictions the most, although they were informed that these predictions would likely be most valid. A Friedman test revealed significant differences in mean ranks between the procedures ($\chi^2(4) = 83.781$, $p < .001$). To determine whether participants significantly preferred the adjustment procedure, we compared it to the general weights procedure (second smallest rank) using a Wilcoxon signed rank test. This difference in mean ranks was moderate to large ($z = -5.77$, $p < .001$, $r = -.47$, 95% CI [−.59, −.33]). So, when participants had to choose a procedure for making valid future admission decisions, they clearly preferred adjusting optimal model predictions.

6.5 | Use of autonomy-enhancing features

The results in the supporting information (Section S2) show that almost all participants used the autonomy-enhancing features. Furthermore, we investigated how participants weighted the predictors, by regressing their predictions on the predictors and subsequently calculating relative weights (Tonidandel & LeBreton, 2015). Figure S1 in the supporting information shows relative weights for the holistic, adjustment, and optimal model condition and the averaged weights in the general weights condition. Compared with the optimal model, participants weighted the personal statement more heavily in all conditions, and the difference was larger when predicting chance of dropout.

6.6 | Predictive validity

For each condition, we computed the correlation coefficient between the predicted scores and the two outcomes (first-year GPA and drop out). We also computed a multiple correlation coefficient based on an optimal regression model for each condition. Because the applicants that were judged differed per condition due to random allocation of applicants to participants, the multiple R based on an optimal regression model differed slightly between conditions. Therefore, we always compared the validity of condition-specific participants' predictions with the validity of a condition-specific optimal regression model.[5] We compared correlations by computing posterior distributions of differences between correlations using the BayesFactor package (0.9.12–4.2), and we reported the 95% HDIs.[6]

6.7 | First-year GPA

The correlation coefficients are presented in Figure 2. The predictive validity of participants' predictions was similar across conditions. However, as expected, the optimal model resulted in the highest predictive validity in each condition ($\bar{r} = .50$).[7] Contrary to expectations, the holistic predictions did not result in the lowest predictive validity ($r = .43$). The holistic predictions had higher validity than predictions made based on individual weights ($r = .36$) and similar validity to predictions made based on general weights ($r = .41$) and holistically adjusted model predictions ($r = .39$). Only the optimal model ($r = .54$) resulted in higher predictive validity than holistic predictions. Other conditions than the optimal condition resulted in slightly lower validity than the holistic condition. These differences were in the opposite direction of our hypothesis.

TABLE 3 Mean rank per condition and rank frequencies in Study 1 and Study 2

Study	Condition	Mean rank	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Study 1	Holistic	3.53	15	24	27	34	50
	Individual	3.13	15	34	37	44	20
	General	3.11	16	31	47	32	24
	Adjust	1.99	84	21	15	23	7
	Optimal	3.23	20	40	24	17	49
Study 2	Holistic	3.75	21	16	35	38	82
	General	3.14	24	38	43	62	25
	General averaged	2.70	36	50	59	29	18
	Adjust	2.15	84	45	25	27	11
	Optimal	3.27	27	43	30	36	56

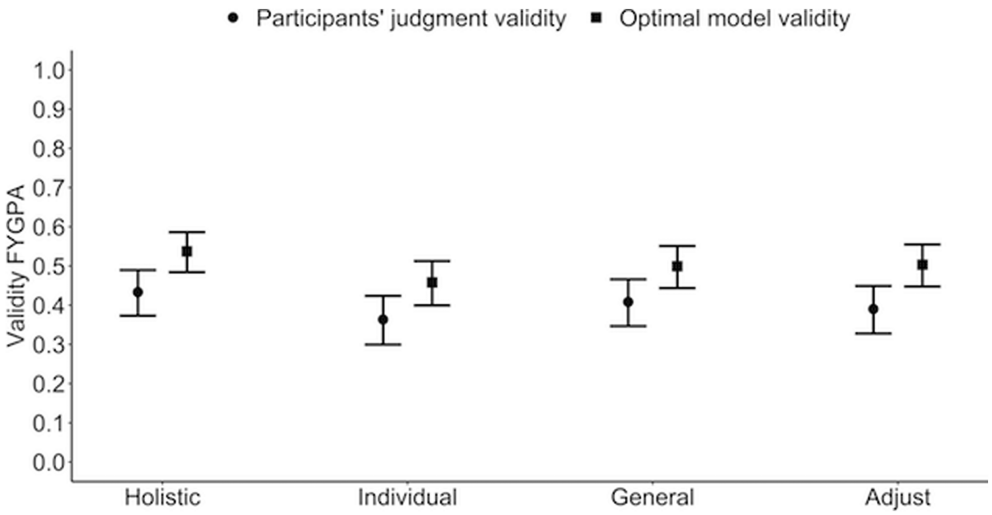

FIGURE 2 Validity for the prediction of first-year GPA in Study 1. *Note*: Error bars represent 95% confidence intervals

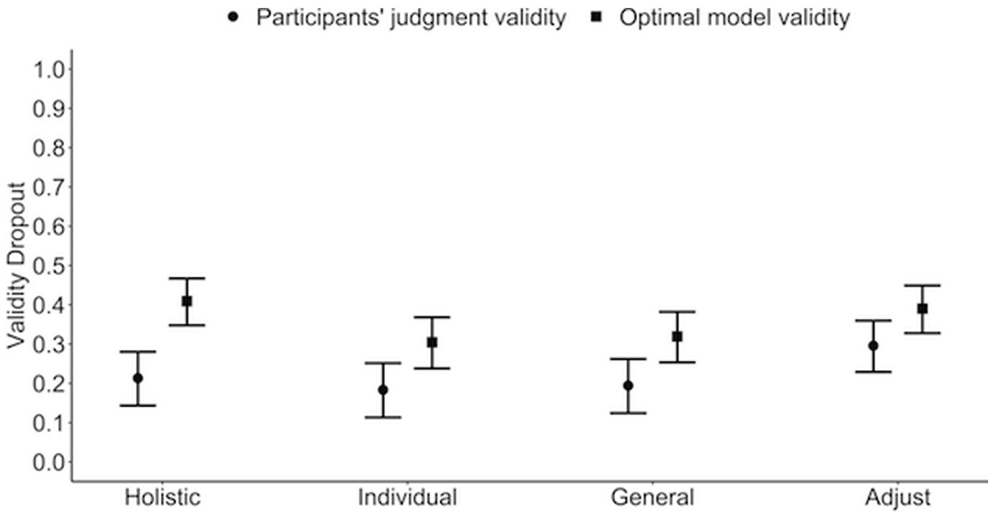

FIGURE 3 Validity for the prediction of dropout in Study 1. *Note*: Error bars represent 95% confidence intervals

6.8 | Dropout

The results for predicting dropout mostly aligned with those for predicting first-year GPA (see Figure 3). The holistic predictions did not result in the lowest predictive validity ($r = .21$) but in validity similar to those based on individual weights ($r = .18$) and general weights ($r = .19$). Again, the optimal model had by far the largest predictive validity ($\bar{r} = .39$).[8] Furthermore, adjusted optimal model predictions ($r = .30$) also resulted in higher validity than holistic predictions ($r_{diff} = -.11$, 95% HDI [−.20, −.03]). So, of all autonomy-enhancing procedures, only holistically adjusted optimal model predictions were more valid than holistic predictions.

6.9 | Exploratory analyses

In the general weights condition, weights were only consistent for the five applicants that were judged by a single participant. However, each participant chose different weights, which reduced consistency within the entire sample of predictions. We also averaged all participants' weights (see Figure S1 in the supporting information) to check whether full consistency would increase predictive validity. Averaging weights barely increased the validity for first-year GPA (from $r = .41$ to $r = .42$) and dropout (from $r = .19$ to $r = .22$). As Figure S1 also shows, there was relatively little variance in the weights that participants chose, which explains why averaged weights did not increase predictive validity much.

7 | DISCUSSION

The results of Study 1 showed that, contrary to expectations, differences in perceived autonomy between all autonomy-enhancing prediction procedures were negligible, as were differences between making holistic predictions and all autonomy-enhancing prediction procedures. Yet, as predicted, decision makers experienced much

more autonomy when making holistic predictions and when using autonomy-enhancing prediction procedures than when using prescribed optimal model predictions. This suggests that participants were relatively insensitive to *how* autonomy was enhanced (by choosing individual/general weights or by adjusting optimal model predictions). So, Hypothesis 1 was mostly unsupported. The results were different for use intentions. Although holistic prediction is predominantly used in practice, participants unexpectedly preferred choosing general predictor weights and adjusting optimal model predictions over all other prediction procedures. So, Hypothesis 2 was also mostly unsupported. One explanation may be that these two autonomy-enhancing prediction procedures require less effort than making holistic predictions (Nolan & Highhouse, 2014), which may also explain why participants disliked choosing weights for each specific applicant. Yet the optimal model condition, which required the least effort, was also disliked the most.

Choosing one set of predictor weights and holistically adjusting optimal model predictions resulted in high use intentions. However, contrary to expectations, predictive validity barely increased compared with holistic predictions. For predicting first-year GPA, choosing predictor weights and adjusting optimal model predictions resulted in very similar predictive validity compared with holistic predictions. For predicting dropout, only adjusting optimal model predictions resulted in slightly higher predictive validity than holistic predictions. Furthermore, our finding that letting participants choose one set of predictor weights did not result in higher validity than holistic predictions cannot be explained by a lack of consistency between participants (Yu & Kuncel, 2020), as averaging weights across participants barely increased predictive validity. An alternative explanation may be that participants weighted the personal statement more heavily when choosing general weights than when making holistic predictions (as shown in Figure S1 in the supporting information). In general, our findings are only partially in line with results presented by Dietvorst et al. (2018), who found that holistically adjusting optimal model predictions resulted in higher validity than fully holistic predictions.

8 | STUDY 2

The aim of Study 2 was to answer a number of questions that were raised by Study 1. In Study 1, holistic predictions were surprisingly valid. This may have been a result of providing participants with predictor validity information, which generally increases validity (Balzer et al., 1989), likely due to increased consistency. However, in practice, predictor validity information may be unavailable or less salient than it was for the participants in Study 1. Therefore, we randomly varied between-subjects whether participants received predictor validity information (presented in the same way as in Study 1) or not, expecting that participants would make more valid predictions when validity information is available.

Dietvorst et al. (2018) found that decision makers used mechanical rule predictions as long as they could adjust them holistically, regardless of the extent of allowed adjustments. Therefore, to improve predictive validity in the adjustment condition, we restricted the extent to which participants could adjust optimal model predictions. Participants could adjust model predictions of first-year GPA by 0.5 grade points (upward or downward) and model predictions of the chance of dropout by 5%. Restricting participants' adjustment behavior allowed us, together with the results from Study 1, to formulate more specific hypotheses. We expected that participants would perceive less autonomy when restrictedly adjusting optimal model predictions compared with choosing predictor weights (cf. Hypothesis 6). Yet, based on the results presented by Dietvorst et al. (2018), we expected that use intentions would be similar to Study 1, where adjustment was unrestricted. Furthermore, we expected that restrictedly adjusting optimal model predictions would result in more valid predictions than freely choosing consistent predictor weights without valid reference points (cf. Hypothesis 11).

A couple of changes were made to investigate attitudes towards prediction procedures more elaborately. At the end of the general weights condition, participants completed the attitudinal measures twice, assuming that (1) their own weights would be used and (2) their weights would be averaged with the weights from other participants. This was done to explore how decision makers would perceive a procedure in which solely their own weights are used, compared with a procedure where their weights are averaged with weights of other decision makers. Participants could also choose this "averaged weights procedure" when rank-ordering the prediction procedures according to which procedure they would most likely use. We also asked participants to report how valid they thought each predictor was. This was done because decision makers' beliefs in predictor validities do not always align with empirical validities (Rynes et al., 2002). However, such differences between beliefs and empirical validities are important to investigate because beliefs may partly determine how decision makers weight predictors (Kuncel, 2018).

In contrast to Study 1, Study 2 was conducted online, and each condition included 10 predictions and a prediction trial to familiarize participants with the task before their prediction was rewarded (as in Study 1, participants could earn a reward up to €5). Furthermore, to reduce the duration of the study, we dropped the holistic predictions at the end of the general weights condition as well as the individual weights condition, which was disliked by participants and resulted in low validity. Moreover, in the optimal model condition, we actually *showed* participants the regression model predictions that they could not adjust. This allowed us to counterbalance all conditions.[9] The median time it took to complete the study was 64 min.

Hypothesis 6. Perceived autonomy will decrease from the holistic condition (most autonomy), general weights condition, and adjustment condition to the optimal model condition (least autonomy).

Hypothesis 7. Participants will report higher use intentions in the general weights condition and the adjustment condition than in the optimal model condition and

higher use intentions in the holistic condition than in the optimal model condition.

We had the same hypotheses for confidence (H8) and satisfaction (H9).

Hypothesis 10. Participants who receive predictor validity information will make predictions that result in higher predictive validity in each condition (except for the optimal model condition) than participants who do not receive predictor validity information.

Hypothesis 11. Predictive validity will increase from the holistic condition (lowest predictive validity), general weights condition, and adjustment condition to the optimal model condition.

9 | METHOD

9.1 | Participants

We conducted power analyses for the repeated measures ANOVAs (Hypotheses 6–9) and for the difference tests between two independent correlations (Hypothesis 10). The latter power analysis required the highest sample size ($N = 192$), assuming 1-β > .8, $\alpha = .05$, allocation ratio N2/N1 = 1, predictions per condition = 10, and an effect size of $q = 0.11$ based on the results from Study 1. We recruited participants in two ways. Participants recruited via the research platform of the University of Groningen received €10 compensation for their participation. Participants from the first-year psychology student pool took part for course credit. In total, 269 participants completed the experiment. No participants from Study 1 participated in Study 2. We excluded participants who failed at least one of three attention checks (see Appendix A) or completed the study in less than 20 min ($n = 77$). The mean age of the final sample ($N = 192$) was $M = 21.69$ ($SD = 6.04$, range 16–64), and the majority of participants was female (77%) and Dutch (93%). All other participants had another European nationality (7%). Furthermore, 88% were enrolled as a student, 11% were employed, and 1% was unemployed. This study was approved by the ethics committee of the Heymans institute for psychological research (code: PSY-1920-S-0120).

9.2 | Measures and stimulus material

We used the exact same attitudinal measures as in Study 1. The scales for use intentions and autonomy showed good reliability ($\alpha = .87$ and $\alpha = .92$). Furthermore, we used a one-item measure to assess how effective participants considered each predictor to be (1 = *not effective*; 5 = *very effective*). As in Study 1, the inter-rater reliability for the personal statement ratings was low ($ICC = .11$). The personal statement rating was a poor predictor of first-year GPA ($\bar{r} = .06$) and dropout ($\bar{r} = -.02$).

10 | RESULTS

10.1 | Attitudinal measures

We ran two-factor mixed-effects Bayesian ANOVAs with the condition as a within-subject factor (fixed effect) and subjects as a random effect, for each of the four dependent measures. The between-subjects factor was not included in the ANOVAs[10] because we only hypothesized differences between the two levels of the between-subjects factor for predictive validity. Figure 4 displays the observed

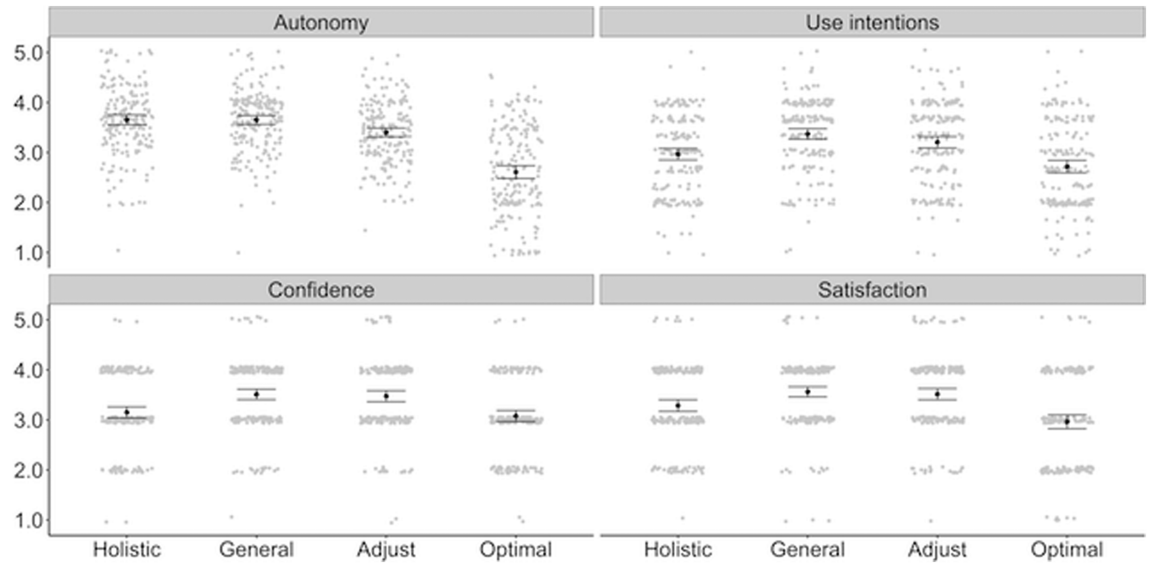

FIGURE 4 Observed means (descriptives) for all conditions per attitude measure in Study 2. *Note*: Error bars represent 95% confidence intervals. Some jittering was added to the plot to improve readability

means per attitudinal measure for each condition. Means and standard deviations for each attitudinal measure in each condition are reported in Table 4 and correlations between attitudinal measures in Table 5. As in Study 1, mean patterns and correlations for the confidence and satisfaction measures were very similar to the mean patterns and correlations of the use intentions measure. Therefore, we report the results for confidence and satisfaction in the supporting information (Section S3).

10.2 | Perceived autonomy

In general, we observed the decreasing trend in perceived autonomy in line with Hypothesis 6. Although perceived autonomy was very similar for making holistic predictions and for choosing predictor weights ($M_{holistic} - M_{general} = 0.002$, 95% HDI [−0.12, 0.12], $d = 0.004$), it was higher for choosing predictor weights than for adjusting optimal model predictions ($M_{general} - M_{adjust} = 0.25$, 95% HDI [0.12, 0.37], $d = 0.41$). This was expected (cf. Hypothesis 6) because we restricted the range in which participants could adjust optimal model predictions but not the range of predictor weights they could choose. Furthermore, as expected, participants perceived much more autonomy when they could adjust optimal model predictions than when they had to use optimal model predictions ($M_{adjust} - M_{optimal} = 0.79$, 95% HDI [0.67, 0.92], $d = 1.03$). So, Hypothesis 6 was mostly supported.

10.3 | Use intentions

As expected, participants' use intentions were much higher for choosing predictor weights than for using optimal model predictions ($M_{general} - M_{optimal} = 0.65$, 95% HDI [0.53, 0.79], $d = 0.81$) and higher for adjusting optimal model predictions than for using optimal model predictions ($M_{adjust} - M_{optimal} = 0.49$, 95% HDI [0.36, 0.62], $d = 0.58$). Furthermore, participants showed higher intentions to use holistic predictions compared with optimal model predictions ($M_{holistic} - M_{optimal} = 0.25$, 95% HDI [0.12, 0.38], $d = 0.29$). Therefore, Hypothesis 7 was fully supported.

10.4 | Procedure ranking

The mean ranks and rank frequencies per procedure are displayed in Table 3. The adjustment procedure was most often (44%) ranked as most preferred for making future admission decisions. In contrast, only 14% of participants preferred to use optimal model predictions the most, although, as in Study 1, they were informed that this procedure would likely be most valid. Also, only a minority (11%) preferred the holistic procedure the most. Interestingly, participants ranked the "averaged weights procedure" first more often than the procedure where solely their own weights were used, although this difference was small (19% vs. 13%). A Friedman test revealed significant differences in mean ranks between the procedures, $\chi^2(4) = 112.83$, $p < .001$. A follow-up Wilcoxon signed rank test showed that participants moderately preferred to adjust optimal model predictions compared with using averaged weights (second smallest rank, $z = -3.70$, $p < .001$, $r = -.27$, 95% CI [−.40, −.13]).

10.5 | Validity beliefs

To investigate whether participants actually believed in the presented predictor validity information, we compared beliefs between participants who did and those who did not receive predictor validity information.[11] Participants who received validity information considered the predictor high school grade to be more effective ($M = 3.97$, $SD = 0.68$) than participants who did not receive validity information

TABLE 4 Means and standard deviations per condition and attitudinal measure in Study 2

	Autonomy		Use intentions		Confidence		Satisfaction	
Condition	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Holistic	3.65	0.70	2.97	0.80	3.15	0.78	3.29	0.82
General	3.65	0.61	3.37	0.73	3.51	0.73	3.56	0.71
Adjust	3.40	0.61	3.20	0.81	3.47	0.77	3.52	0.79
Optimal	2.61	0.90	2.72	0.87	3.08	0.79	2.96	0.97

TABLE 5 Correlations between attitudinal measures autonomy, use intentions, confidence, and satisfaction in Study 2

Measure	Autonomy	Use intentions	Confidence	Satisfaction
Autonomy	-			
Use intentions	.41*	-		
Confidence	.35*	.67*	-	
Satisfaction	.42*	.70*	.73*	-

Note: Correlations were averaged across conditions using Fisher's z transformation. Correlations between attitudinal measures per condition are reported in the supporting information in Table S3.
*$p < .05$.

($M = 3.67$, $SD = 0.74$, $M_{diff} = 0.29$, 95% HDI [0.09, 0.49], $d = 0.41$). Validity beliefs regarding the effectiveness of the admission test score were similar for participants who received validity information ($M = 4.16$, $SD = 0.64$) and those who did not receive validity information ($M = 4.21$, $SD = 0.76$, $M_{diff} = -0.05$, 95% HDI [−0.25, 0.14], $d = -0.08$). Lastly, participants who received validity information considered the personal statement to be much less effective ($M = 2.45$, $SD = 0.92$) than participants who did not receive validity information ($M = 3.39$, $SD = 1.05$, $M_{diff} = -0.92$, 95% HDI [−1.21, −0.63], $d = -0.95$). This shows that predictor validity beliefs were more aligned with actual predictor validities when participants received validity information.

10.6 | Use of autonomy-enhancing features

The results presented in the supporting information (Section S3) show that, as in Study 1, almost all participants made use of the autonomy-enhancing features. Furthermore, the relative weights analysis showed that, compared with the optimal model, participants weighted the personal statement more heavily in all conditions, and the difference was larger when predicting chance of dropout (see Figure S2). Moreover, participants who did not receive validity information weighted the personal statement slightly more heavily than participants who received validity information.

10.7 | Predictive validity

10.7.1 | First-year GPA

Figure 5 displays the validity of participants' predictions for those who did and did not receive validity information separately, and the optimal model validity. In the holistic and adjustment condition, predictive validity was very similar between participants who received validity information and participants who did not receive validity information (holistic: $r = .31$ and $r = .32$; $r_{diff} = -.01$, 95% HDI [−.09, .07] and adjustment: $r = .34$ and $r = .36$; $r_{diff} = -.02$, 95% HDI [−.10, .05], respectively). However, in the general weights condition, participants who received validity information made slightly more accurate predictions than participants who did not receive validity information ($r = .41$ and $r = .34$; $r_{diff} = .07$, 95% HDI [−.01, .15]). So, presenting participants with validity information only increased predictive validity in the general weights condition, albeit only slightly (partial support for Hypothesis 10).

Hypothesis 11 stated that predictive validity increases from the holistic condition to the general weights condition to the adjustment condition to the optimal model condition. For participants who did not receive predictor validity information, predictive validity slightly increased from the holistic condition ($r = .32$) to the general weights condition ($r = .34$) to the adjustment condition ($r = .36$) to the optimal model condition ($\bar{r} = .41$). For participants who received predictor validity information, predictive validity increased markedly from the holistic condition ($r = .31$) to the general weights condition ($r = .41$). However, predictive validity in the adjustment condition ($r = .34$) was not higher than predictive validity in the general weights condition. Furthermore, optimal model predictions were as valid ($\bar{r} = .41$) as predictions in the general weights condition. So, we found some evidence for the predicted increase in predictive validity from the holistic condition to the optimal model condition, although the general weights condition resulted in higher predictive validity than the adjustment condition when participants had predictor validity information (partial support for H11).

10.7.2 | Dropout

Figure 6 displays the validity of participants' predictions and the optimal model validity for predicting dropout. These results are quite different from the results obtained for predicting first-year GPA. Unexpectedly, in the holistic condition, participants who had validity information made *less* accurate predictions than participants without

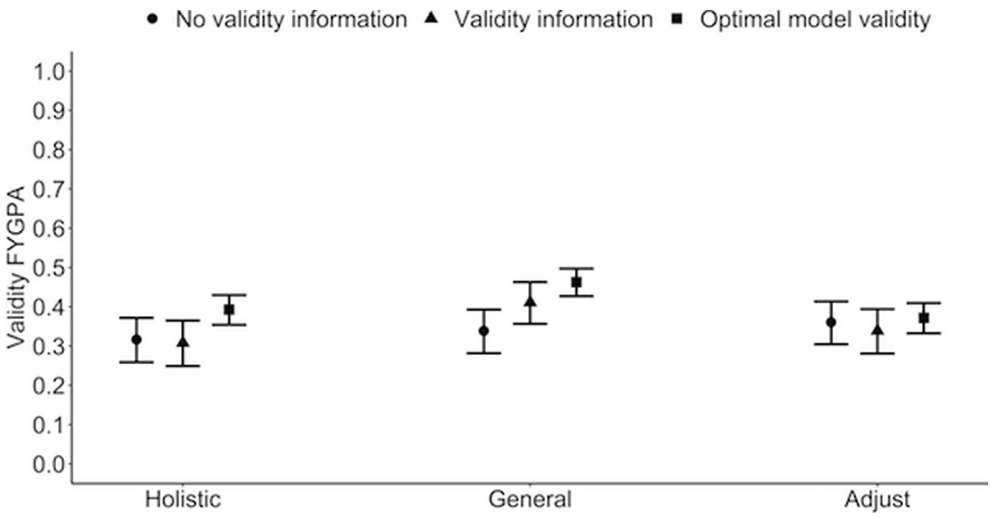

FIGURE 5 Validity for the prediction of first-year GPA in Study 2. *Note*: Error bars represent 95% confidence intervals

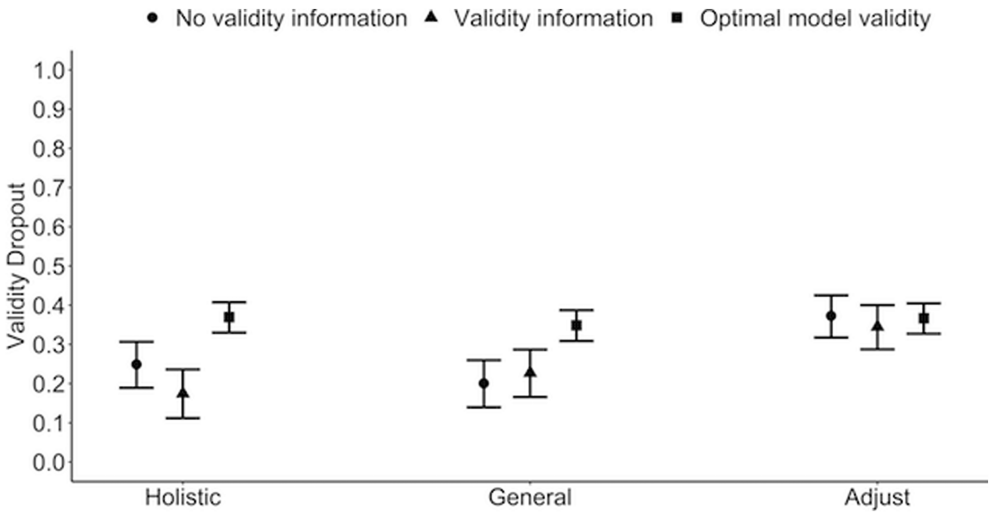

FIGURE 6 Validity for the prediction of dropout in Study 2. *Note*: Error bars represent 95% confidence intervals

validity information ($r = .18$ and $r = .25$; $r_{diff} = -.07$, 95% HDI [−.16, .01]). In the general weights and adjustment condition, participants with validity information made similarly accurate predictions as participants without validity information (general: $r = .23$ and $r = .20$; $r_{diff} = .03$, 95% HDI [−.06, .11] and adjustment: $r = .35$ and $r = .37$; $r_{diff} = -.03$, 95% HDI [−.11, .05], respectively).

For participants who did not receive predictor validity information, predictive validity *decreased* from the holistic condition ($r = .25$) to the general weights condition ($r = .20$). However, the adjustment condition ($r = .37$) resulted in more accurate predictions than the general weights condition. Furthermore, optimal model predictions ($\bar{r} = .36$) were as valid as predictions in the adjustment condition. For participants who received predictor validity information, predictive validity increased from the holistic condition ($r = .17$) to the general weights condition ($r = .23$) to the adjustment condition ($r = .35$). So, support for Hypothesis 11 differed depending on whether validity information was provided. We found support for Hypothesis 11 when validity information was provided but inconsistent results when validity information was not provided.

10.8 | Exploratory analysis

10.8.1 | Averaged general weights condition

If decision makers determine predictor weights in practice, it is likely that multiple decision makers are involved. To resemble this scenario, we also asked participants to report their attitudes towards using weights that are averaged across participants. In contrast to the results from the ranking procedure, these results showed that participants liked the idea to use averaged weights less than using their own weights. Compared with using their own weights, participants experienced less autonomy in an approach where weights would be averaged ($M_{diff} = 0.29$, 95% HDI [0.19, 0.39], $d = 0.45$). We obtained similar results for use intentions ($M_{diff} = 0.20$, 95% HDI [0.09, 0.31], $d = 0.29$), confidence ($M_{diff} = 0.18$, 95% HDI [0.06, 0.29], $d = 0.25$), and satisfaction ($M_{diff} = 0.22$, 95% HDI [0.11, 0.33], $d = 0.32$). Furthermore, using averaged weights did not increase predictive validity compared with using participants' own weights when validity information was present ($r = .41$ and $r = .41$ for first-year GPA and $r = .23$ and $r = .23$ for dropout). When validity information was absent, averaging weights only slightly increased predictive validity (from $r = .34$ to $r = .39$ for first-year GPA and from $r = .20$ to $r = .24$ for dropout). In sum, using averaged weights resulted in more negative attitudes but in very similar predictive validity compared with using participants' own weights.

11 | DISCUSSION

In line with the results from Study 1 and our expectations, participants experienced much more autonomy when making holistic predictions and when using autonomy-enhancing prediction procedures than when using optimal model predictions. Furthermore, as expected, participants perceived less autonomy when restrictedly adjusting optimal model predictions than when freely choosing predictor weights. Like in Study 1, participants were more likely to use autonomy-enhancing prediction procedures compared with using holistic and optimal model predictions. In terms of validity, presenting participants with validity information only increased validity when choosing predictor weights for predicting first-year GPA, albeit only slightly. So, the unexpectedly high validity in the holistic condition does not seem to be fully explained by the availability of validity information. Furthermore, the expectation that validity would increase from the holistic condition, general weights condition, and adjustment condition to the optimal condition (Hypothesis 11) was mostly unsupported; we found this pattern for dropout predictions but only when validity information was available.

While we did not have an a priori hypothesis, our results suggest that it may be somewhat easier for decision makers to translate validity information into explicit predictor weights than to use this information holistically, although earlier research suggests that validity

information also helps when making holistic predictions (Balzer et al., 1989). Our results also showed that predictor validity beliefs were better aligned with actual predictor validities when participants received validity information, which was also reflected in the predictor weights that participants chose (see Figure S2). Interestingly though, even when validity information was present, participants considered the personal statement to be more important for dropout predictions than for first-year GPA predictions. This underscores that, besides criterion-related validity, face validity can play an important role in decision-making (also see Kuncel, 2018, p. 475).

12 | GENERAL DISCUSSION

In two studies, we investigated the effect of several autonomy-enhancing features that participants had when making predictions on perceptions and predictive validity. In contrast to expectations, most of our hypotheses were not supported. Yet, in both studies, perceived autonomy was much higher in all autonomy-enhancing prediction procedures than when using optimal model predictions. In line with earlier findings by Nolan and Highhouse (2014) and Dietvorst et al. (2018), we found clear evidence that participants preferred choosing their own, consistent (but not applicant-specific) predictor weights and adjusting optimal model predictions over holistic and optimal model predictions. So, although perceived autonomy was similar for holistic predictions and autonomy-enhancing prediction procedures, use intentions differed between these conditions, suggesting that other factors besides autonomy affect use intentions.

We found mixed evidence that autonomy-enhancing prediction procedures resulted in higher validity than holistic predictions. In Study 1, only adjusted optimal model predictions of dropout were more valid than holistic predictions. In Study 2, restrictedly adjusting optimal model predictions resulted in slightly higher validity than holistic predictions, as did choosing predictor weights but only for predicting first-year GPA and only when predictor validity information was available. So, our results tentatively suggest that the trade-off between autonomy need satisfaction and validity is optimized when decision makers can restrictedly adjust optimal model predictions or choose consistent predictor weights when validity information is available to decision makers.

In earlier research, choosing predictor weights (Nolan & Highhouse, 2014) and adjusting optimal model predictions (Dietvorst et al., 2018) as a means to overcome algorithm aversion have been investigated in separate studies. We investigated both prediction procedures to allow for a comparison in attitudes and validity. Extending earlier research, we showed that use intentions for these two autonomy-enhancing prediction procedures were similar, although ranking the different prediction procedures resulted in a preference for adjusting optimal model predictions. One reason why these results were different may be that we explicitly asked and incentivized participants to rank the prediction procedures according to predictive validity, while the use intentions scale items were generic and did not explicitly ask participants to evaluate the procedures with regard to predictive validity. Aside from predictive validity, decision makers also consider other aspects of selection procedures, such as costs, effort, and fairness (König et al., 2010; Neumann, Niessen, et al., 2021; Pyburn et al., 2008). It may be that decision makers prefer adjusting optimal model predictions when predictive validity is the most important aspect, because they are confident that they can find valid rule exceptions (Dietvorst et al., 2018). When also considering other aspects, choosing predictor weights may be preferred because this procedure does not require judging each applicant individually and treats each applicant equally (Nolan et al., 2016).

Relatedly, results based on the use intentions scale suggested that participants preferred solely using their own weights over the averaged weights procedure. In contrast, results based on the ranking procedure showed that participants had a small preference for the averaged weights procedure over solely using their own weights. This suggests that decision makers may trust "the wisdom of the crowd" when predictive validity is most important. Another explanation for these findings may be that the joint evaluation mode in the ranking task allowed participants to differentiate more easily between prediction procedures than the separate evaluation mode of the use intentions measure (Hsee & Zhang, 2010).

12.1 | Limitations and future research

Compared with our student samples, experienced professionals may evaluate prediction procedures such as choosing predictor weights and adjusting optimal model predictions more negatively, as they may feel less need for prediction support (Arkes et al., 1986; Dawes, 1976). Therefore, a replication with experienced professionals such admission officers, HR professionals, or clinical psychologists would be very valuable.

Our choice to use a within-subjects design had advantages and disadvantages. Besides increasing power, it allowed for a joint evaluation of prediction procedures (Nolan et al., 2020). This is most representative of decision-making in practice when deciding on what procedure to adopt. Yet, it is plausible that smaller effect sizes would have been observed for our attitudinal measures in a between-subjects design (i.e., in separate evaluation) because prediction procedures could not be easily compared. Furthermore, the within-subjects design required us to vary the stimuli (applicants) between conditions, which resulted in small variations in the optimal model validity. Therefore, validity differences between prediction procedures could be partly due to sampling error variance. Another consequence of our within-subjects design may have been that participants applied the prediction procedure they experienced first also in later conditions, although our results in the supporting information suggest that the condition order had a negligible effect on participants' behavior. In any case, it should be noted that earlier studies utilized a between-subjects design, which may explain why our results were only partially in line with the results found by Dietvorst et al. (2018). For these

reasons, future studies could use a between-subjects design and the same applicants across conditions. Lastly, in the adjustment condition, participants knew that the regression model predictions did not include the personal statement. Although adding the personal statement ratings as a predictor would have barely changed the model predictions, participants may have adjusted model predictions less if this predictor was included.

In our studies, we labeled predictors as "good" or "bad" for ease of communication. These descriptions were rather unspecific and may explain why decision makers still assigned a nontrivial weight to the personal statement. Therefore, validity information may be presented more specifically and in different formats, such as in the form of suggested percentage predictor weights or even in a narrative manner. Alternatively, in future studies, decision makers' choice in predictor weights may be restricted to avoid that poor predictors receive a substantial weight. However, restricting the choice of weights or the adjustment of rule predictions requires someone to set these restrictions and to control whether decision makers adhere to it, which may be difficult in practice.

Researchers are also encouraged to investigate predictive validity of and decision makers' attitudes towards a prediction procedure that combines autonomy in the design of a mechanical rule and in the adjustment of the rule's result. Based on cognitive dissonance theory (Festinger, 1957), it may be expected that decision makers are less likely to adjust a rule's prediction when they have designed the rule themselves.

Moreover, the nature of predictors may influence the effort that is required to process information, which could affect intentions to use certain prediction procedures and how decision makers weight information when making predictions. Therefore, the nature of predictors could be varied in future studies. For example, predictors could be presented quantitatively (e.g., test scores or interview ratings) or qualitatively (e.g., recordings of an interview, written personal statements). Also, researchers may investigate other selection contexts than admission to higher education and vary the manner in which the criteria are predicted. For example, in personnel selection, decision makers may predict applicants' job performance by classifying them into more general categories (e.g., red, yellow, and green; see Hoffman et al., 2017). When decision makers are restricted in adjusting optimal model predictions, they may experience restrictions on shifts between categories as more consequential than on adjustments to continuous performance predictions.

Lastly, researchers may investigate how alternative prediction procedures satisfy decision makers' competence and relatedness needs. There exists some evidence that competence needs are less satisfied in mechanical prediction, compared with holistic prediction (Nolan, 2013) because decision makers cannot demonstrate their assumed expertise in combining information (Nolan et al., 2016). Relatedness needs may be reduced in mechanical prediction because holistic prediction often takes place in group settings (Bolander & Sandberg, 2013). In future research, it could be investigated how relatedness needs are satisfied when decision makers design a mechanical rule in a group setting rather than individually as in our studies. Relatedly, it should be investigated whether stakeholders give more credit to decision makers who use a self-designed rather than an existing mechanical rule (Nolan et al., 2016, 2020).

12.2 | Practical recommendations

Our results suggest that, compared with holistic prediction, slightly more valid predictions were made when decision makers could restrictedly adjust optimal model predictions. This approach requires that an optimal rule is available, which is not always the case in practice. Alternatively, decision makers could choose a consistent set of predictor weights themselves. Yet, when predictors substantially differ in predictive validity, which is often the case in practice, care should be taken that decision makers choose appropriate weights. Our results suggest that one way to support decision makers in designing their own mechanical rules effectively is to provide them with predictor validity information.

12.3 | Conclusion

One reason for the underutilization of mechanical prediction is that it restricts decision makers' autonomy in the decision-making process. We contribute to the literature on interventions to promote the use of mechanical prediction in practice (Dietvorst et al., 2018; Neumann, Niessen, et al., 2021) by showing that decision makers are much more likely to use mechanical prediction procedures when their autonomy is enhanced by the option to choose their own, consistent predictor weights, or to adjust mechanical rule predictions. However, contrary to expectations, we found inconclusive evidence that these autonomy-enhancing mechanical prediction procedures resulted in higher predictive validity than holistic predictions, and the increase in predictive validity was rather small. To close the science-practice gap in decision-making, more research is needed to identify ways to increase the validity of autonomy-enhancing mechanical prediction procedures.

DATA AVAILABILITY STATEMENT

The data that we are allowed to share that support the findings of the studies reported in this paper are openly available on OSF at https://osf.io/86jfb/.

ORCID

Marvin Neumann https://orcid.org/0000-0003-0193-8159
A. Susan M. Niessen https://orcid.org/0000-0001-8249-9295
Jorge N. Tendeiro https://orcid.org/0000-0003-1660-3642
Rob R. Meijer https://orcid.org/0000-0001-5368-992X

ENDNOTES

[1] Although we used Bayesian parameter estimation to analyze our results, we reported the results from our preregistered frequentist power analyses so that readers who are mostly familiar with frequentist

testing can understand that our studies had sufficient power to detect significant differences if we had tested our hypotheses using frequentist null hypothesis significance tests.

[2] We also conducted a sensitivity power analysis for a one-tailed dependent *t*-test, given that pairwise comparisons were most important for answering our hypotheses. Given $N = 150$, $\alpha = .05$, and $1\text{-}\beta = .80$, the smallest effect size we could detect with sufficient power was $d_z = 0.20$. Furthermore, for predictive validity, we also calculated the required effect size to detect a significant difference between dependent correlations, assuming $1\text{-}\beta > .8$; $\alpha = .05$, and $N = 750$. The N resulted from 150 participants who each made five judgments in a condition. Given a correlation coefficient of $r_{ab} = .4$, the alternative correlation coefficient must at least be $r_{ac} = .45$, assuming $r_{bc} = 0.8$.

[3] The model did not include the personal statement because ratings were not available before data collection. However, personal statements have very low predictive validity (Murphy et al., 2009). Therefore, we expected that personal statement ratings would not significantly improve a model that contained high school GPA and admission test scores, as indeed was the case (first-year GPA, $\Delta R^2 < .001$, $F(1,746) = .18$, $p = .671$; dropout, Nagelkerke's pseudo R^2, $\Delta R^2_p = .004$, $\chi^2(1) = 1.85$, $p = .173$).

[4] We checked for order effects for predictive validity. The order of conditions had a negligible effect on participants' behavior and differences in predictive validity were in the opposite direction of expectations. We present the results in the supporting information (Section S2).

[5] Cross-validated multiple Rs of the optimal regression models are reported in the supporting information because they barely deviated from the original multiple Rs.

[6] We had to deviate from our preregistration here due to an oversight and misunderstanding among the authors with regard to the preregistered analysis plan.

[7] Correlational differences and 95% HDIs between the validity of participants' predictions and optimal model predictions per condition for first-year GPA: Holistic—optimal = −.10, 95% HDI [−.18, −.03]; Individual—optimal = −.09, 95% HDI [−.18, −.01]; General—optimal = −.13, 95% HDI [−.21, −.05]; Adjust—optimal = −.11, 95% HDI [−.20, −.04].

[8] Correlational differences and 95% HDIs between the validity of participants' predictions and optimal model predictions per condition for dropout: Holistic—optimal = −.20, 95% HDI [−.28, −.10]; Individual—optimal = −.12, 95% HDI [−.22, −.03]; General—optimal = −.12, 95% HDI [−.22, −.03]; Adjust—optimal = −.09, 95% HDI [−.18, −.01].

[9] As in Study 1, the order of conditions had a negligible effect on participants' behavior and differences in predictive validity were in the opposite direction of expectations. We present the results in the supporting information (Section S3).

[10] We also ran the ANOVAs with the between-subjects factor included. The main effect for the between-subjects factor and the interaction effect showed negligible effect sizes (η^2_ps ≤ .014).

[11] We had preregistered that we would also provide Bayes factors for differences in validity beliefs. For simplicity and consistency reasons, we decided to stick to Bayesian parameter estimation throughout the paper and therefore reported mean differences and 95% highest density intervals.

REFERENCES

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Arkes, H. R. (2008). Being an advocate for linear models of judgment is not an easy life. In J. I. Krueger (Ed.), *Rationality and Social Responsibility: Essays in Honor of Robyn Mason Dawes* (pp. 47–70). Psychology Press.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*(1), 93–110. https://doi.org/10.1016/0749-5978(86)90046-4

Balzer, W. K., Doherty, M. E., & O'Connor, R. J. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, *106*(3), 410–433. https://doi.org/10.1037/0033-2909.106.3.410

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1–2), 215–227. https://doi.org/10.1016/S0378-3758(99)00044-0

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods*, *10*(4), 689–709. https://doi.org/10.1177/1094428106294734

Bolander, P., & Sandberg, J. (2013). How employee selection decisions are made in practice. *Organization Studies*, *34*(3), 285–311. https://doi.org/10.1177/0170840612464757

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dana, J., & Thomas, R. (2006). In defense of clinical judgment and mechanical prediction. *Journal of Behavioral Decision Making*, *19*, 413–428. https://doi.org/10.1002/bdm.537

Davis, K. M., Doll, J. F., & Sterner, W. R. (2018). The importance of personal statements in counselor education and psychology doctoral program applications. *Teaching of Psychology*, *45*(3), 256–263. https://doi.org/10.1177/0098628318779273

Dawes, R. M. (1976). Shallow psychology. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and Social Behavior* (pp. 3–11). Lawrence Erlbaum.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106. https://doi.org/10.1037/h0037613

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Deci, E. L., Ryan, R. M., Gagné, M., Leone, D. R., Usunov, J., & Kornazheva, B. P. (2001). Need satisfaction, motivation, and well-being in the work organizations of a former eastern bloc country: A cross-cultural study of self-determination. *Personality and Social Psychology Bulletin*, *27*(8), 930–942. https://doi.org/10.1177/0146167201278002

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314. https://doi.org/10.1177/0956797620948841

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making*, *25*(5), 458–468. https://doi.org/10.1002/bdm.741

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Ganzach, Y., Kluger, A. N., & Klayman, N. (2000). Making decisions from an interview: Expert measurement and mechanical combination. *Personnel Psychology*, *53*(1), 1–20. https://doi.org/10.1111/j.1744-6570.2000.tb00191.x

Gelman, A. (2016). Bayesian inference completely solves the multiple comparisons problem. Statistical modeling, causal inference, and social science. https://statmodeling.stat.columbia.edu/2016/08/22/bayesian-inference-completely-solves-the-multiple-comparisons-problem/

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

GlenMaye, L., & Oakes, M. (2002). Assessing suitability of MSW applicants through objective scoring of personal statements. *Journal of Social Work Education*, *38*(1), 67–82. https://doi.org/10.1080/10437797.2002.10779083

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Guay, J. P., & Parent, G. (2018). Broken legs, clinical overrides, and recidivism risk: An analysis of decisions to adjust risk levels with the LS/CMI. *Criminal Justice and Behavior*, *45*(1), 82–100. https://doi.org/10.1177/0093854817719482

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, *21*(1), 1–21. https://doi.org/10.1037/a0014421

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*(3), 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*(4), 383–389. https://doi.org/10.1111/ijsa.12192

Hoffman, M., Kahn, L. B., & Li, D. (2017). Discretion in hiring. *The Quarterly Journal of Economics*, *133*(2), 765–800. https://doi.org/10.1093/qje/qjx042

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology*, *86*(5), 680–695. https://doi.org/10.1037/0022-3514.86.5.680

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science*, *5*(4), 343–355. https://doi.org/10.1177/1745691610374586

Kaplan, S. E., Reneau, J. H., & Whitecotton, S. (2001). The effects of predictive ability information, locus of control, and decision maker involvement on decision aid reliance. *Journal of Behavioral Decision Making*, *14*(1), 35–50. 10.1002/1099-0771(200101)14:1<35::AID-BDM364>3.0.CO;2-D

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

Kirch, D. G. (2012). Transforming admissions: The gateway to medicine. *JAMA: Journal of the American Medical Association*, *308*(21), 2250–2251. https://doi.org/10.1001/jama.2012.74126

König, C. J., Klehe, U., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment*, *18*(1), 17–27. https://doi.org/10.1111/j.1468-2389.2010.00485.x

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and STAN* (2nd ed.). Academic Press. https://doi.org/10.1016/B978-0-12-405888-0.09999-2

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752. https://doi.org/10.1177/1094428112457829

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

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 and Organizational Psychology* (2nd ed.) (pp. 474–487). SAGE Publications Ltd.. https://doi.org/10.4135/9781473914940

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. https://doi.org/10.1037/a0034156

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

MacDonald, W. (2020). OPINION: Tests are not the problem; how they are used can be. https://hechingerreport.org/opinion-tests-are-not-the-problem-how-they-are-used-can-be/

Meehl, P. E. (1954a). Empirical comparisons of clinical and actuarial prediction. In P. E. Meehl (Ed.), *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (pp. 83–128). University of Minnesota Press. doi:10.1037/11281-008

Meehl, P. E. (1954b). The special powers of the clinician. In *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (pp. 24–28). University of Minnesota Press. https://doi.org/10.1037/11281-004

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E.-J. (2016). Continued misinterpretation of confidence intervals: Response to Miller and Ulrich. *Psychonomic Bulletin & Review*, *23*, 131–140. https://doi.org/10.3758/s13423-015-0955-8

Murphy, K. R., Deckert, P. J., Kinney, T. B., & Kung, M.-C. C. (2013). Subject matter expert judgments regarding the relative importance of competencies are not useful for choosing the test batteries that best predict performance. *International Journal of Selection and Assessment*, *21*(4), 419–429. https://doi.org/10.1111/ijsa.12051

Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. *College and University*, *84*(4), 83–86.

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*. Advance online publication. 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*(3), 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. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLoS ONE*, *13*, Article e0198746. https://doi.org/10.1371/journal.pone.0198746

Nolan, K. P. (2013). Basic psychological need fulfillment and user resistance to objective and analytical decision-making practices in employee selection (doctoral dissertation). http://rave.ohiolink.edu/etdc/view?acc_num=bgsu1343479006

Nolan, K. P., Carter, N. T., & Dalal, D. K. (2016). Threat of technological unemployment: Are hiring managers discounted for using standardized employee selection practices? *Personnel Assessment and Decisions*, *2*(1), 30–47. https://doi.org/10.25035/pad.2016.004

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*(2), 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*(4), 328–346. https://doi.org/10.1080/08959285.2014.929691

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Pyburn, K. M. J., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology*, *61*(1), 143–151. https://doi.org/10.1111/j.1744-6570.2008.00108.x

Ryan, A. M., Inceoglu, I., Bartram, D., Golubovich, J., Grand, J., Reeder, M., Derous, E., Nikolaou, I., & Yao, X. (2015). Trends in testing: Highlights of a global survey. In I. Nikolaou & J. Oostrom (Eds.), *Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice* (pp. 136–153). Routledge/Taylor & Francis Group.

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology*, *40*(3), 455–488. https://doi.org/10.1111/j.1744-6570.1987.tb00610.x

Rynes, S. L. (2012). The research-practice gap in I/O psychology and related fields: Challenges and potential solutions. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology* (Vol. 1) (pp. 409–452). Oxford University Press.

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*(2), 149–174. https://doi.org/10.1002/hrm.10029

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology*, *59*, 419–450. https://doi.org/10.1146/annurev.psych.59.103006.093716

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology*, *48*(5), 593–602. https://doi.org/10.1086/219248

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*(3), 178–200. https://doi.org/10.1037/h0023624

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

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. https://doi.org/10.1002/bdm.486

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice*, *40*(5), 525–531. https://doi.org/10.1037/a0014693

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05.". *American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Yu, M. C., & Kuncel, N. R. (2020). Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions*, *6*(2), 1–10. https://scholarworks.bgsu.edu/pad/vol6/iss2/2, https://doi.org/10.25035/pad.2020.02.002

AUTHOR BIOGRAPHIES

Marvin Neumann is a PhD candidate in the department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, at the University of Groningen, the Netherlands. His research focuses on personnel and educational selection, mechanical prediction, and the scientist-practitioner gap in assessment and selection.

A. Susan M. Niessen is an assistant professor in the department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, at the University of Groningen, the Netherlands. She has published in the areas of personnel and educational selection, test-use and decision-making, predictive validity, applicant perceptions, and test bias.

Jorge Tendeiro is a professor at Hiroshima University. His research interests revolve around item response theory and Bayesian statistics. He is interested in both the theoretical as well as the empirical approaches to statistics and data analysis. He is also an advocate of Open Science.

Rob R. Meijer is a full professor in the Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, the Netherlands. His research focuses on applied psychometrics, educational and personnel selection, and decision-making through tests.

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: Neumann, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2022). The autonomy-validity dilemma in mechanical prediction procedures: The quest for a compromise. *Journal of Behavioral Decision Making*, *35*(4), e2270. https://doi.org/10.1002/bdm.2270

APPENDIX A. ATTENTION CHECKS

A.1.1 | Study 1

Participants answered two attention checks. After reading the study instructions, the first attention check required participants to indicate the maximum reward they could earn in the study. The second attention check required participants to indicate for each predictor (high school GPA, admission test score, and personal statement) whether it is a good or bad predictor of study success.

A.1.2 | Study 2

Participants answered three attention checks. After reading the study instructions, the first attention check required participants to indicate the maximum reward they could earn in the study. For participants who received validity information, the second attention check was the same as in Study 1. Participants who did not receive validity information had to choose which predictor they would not use for making predictions (answer options: mean high school GPA, admission test score, personal statement, and sex). The third attention check was administered before participants ranked the prediction procedures and required them to answer by how much their reward earned up to that point would increase if they chose the most valid procedure (correct answer: 20%).

APPENDIX B. Instructions to Participants

TABLE B1 Instructions given to participants in Study 1 per condition (translated from Dutch to English)

Condition	Instructions
Holistic	In this condition, we first ask you to rate a student's personal statement. After you have given your rating, you will see a student's score on the admission test, high school GPA, and your personal statement rating.
Individual weights	In this condition, we first ask you to rate a student's personal statement. After you have given your rating, you will see a student's admission test score, high school GPA, and your personal statement rating. Based on this information, you indicate how you would like to weight the admission test score, high school GPA, and personal statement rating to predict a student's study success. The higher the weight that you assign, the more the information determines your prediction. If you, for example, assign a weight of 33.3% to the admission test score, high school GPA, and the personal statement rating, then your prediction of a student's first-year GPA will be determined with 33.3% by the admission test score, the high school GPA, and the personal statement rating, respectively. In this condition, you assign weights to the information for each of the five students separately.
General weights	In this condition, we first ask you to indicate how you would like to weight the information for making predictions. We would like you to indicate how you want to weight the admission test score, high school GPA, and personal statement rating to predict a student's study success. The higher the weight that you assign, the more the information determines your prediction. If you, for example, assign a weight of 33.3% to the admission test score, high school GPA, and the personal statement rating, then your prediction of a student's first-year GPA will be determined with 33.3% by the admission test score, the high school GPA, and the personal statement rating, respectively. In this condition, you assign weights to the information, which will apply to all five students that you evaluate in this condition. After you have done this, you will answer some questions about your thoughts and opinion with regard to this prediction procedure. Afterwards, you will rate a student's personal statement, after which you will make predictions for each of the five students separately.
Adjustment	In this condition, we first ask you to rate a student's personal statement. After you have given your rating, you will see a prediction of the student's study success based on a statistical model. The statistical model is based on the student's high school GPA and admission test score and applies weights to this information such that the prediction based on this information is as optimal as possible. The personal statement is not taken into account. The statistical model is quite accurate in predicting a student's first-year GPA and chance of dropout in the psychology bachelor program. The statistical model most likely gives the best prediction possible, better than a human prediction. The prediction based on this model is not perfect though. You can adjust the model prediction in case you want to do so. If you do not want to adjust the model prediction, you simply indicate the model prediction.
Optimal model	Imagine that your predictions would be solely determined by the statistical model and that you could not adjust the model predictions.