VU VRIJE UNIVERSITEIT AMSTERDAM

FACULTY OF BEHAVIOURAL AND MOVEMENT SCIENCES
DEPARTMENT OF EXPERIMENTAL AND APPLIED
PSYCHOLOGY

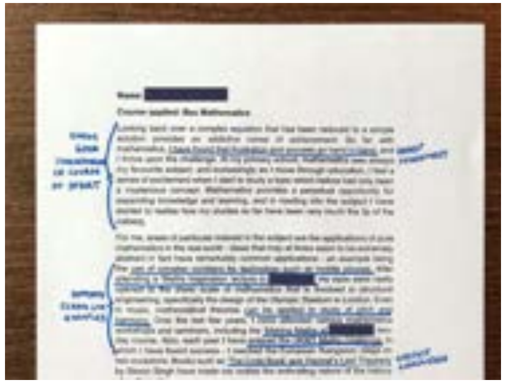
Improving Advisor Systems: Algorithms that Explain Themselves

Nathan R. Kuncel, University of Minnesota

Marvin Neumann, VU University Amsterdam

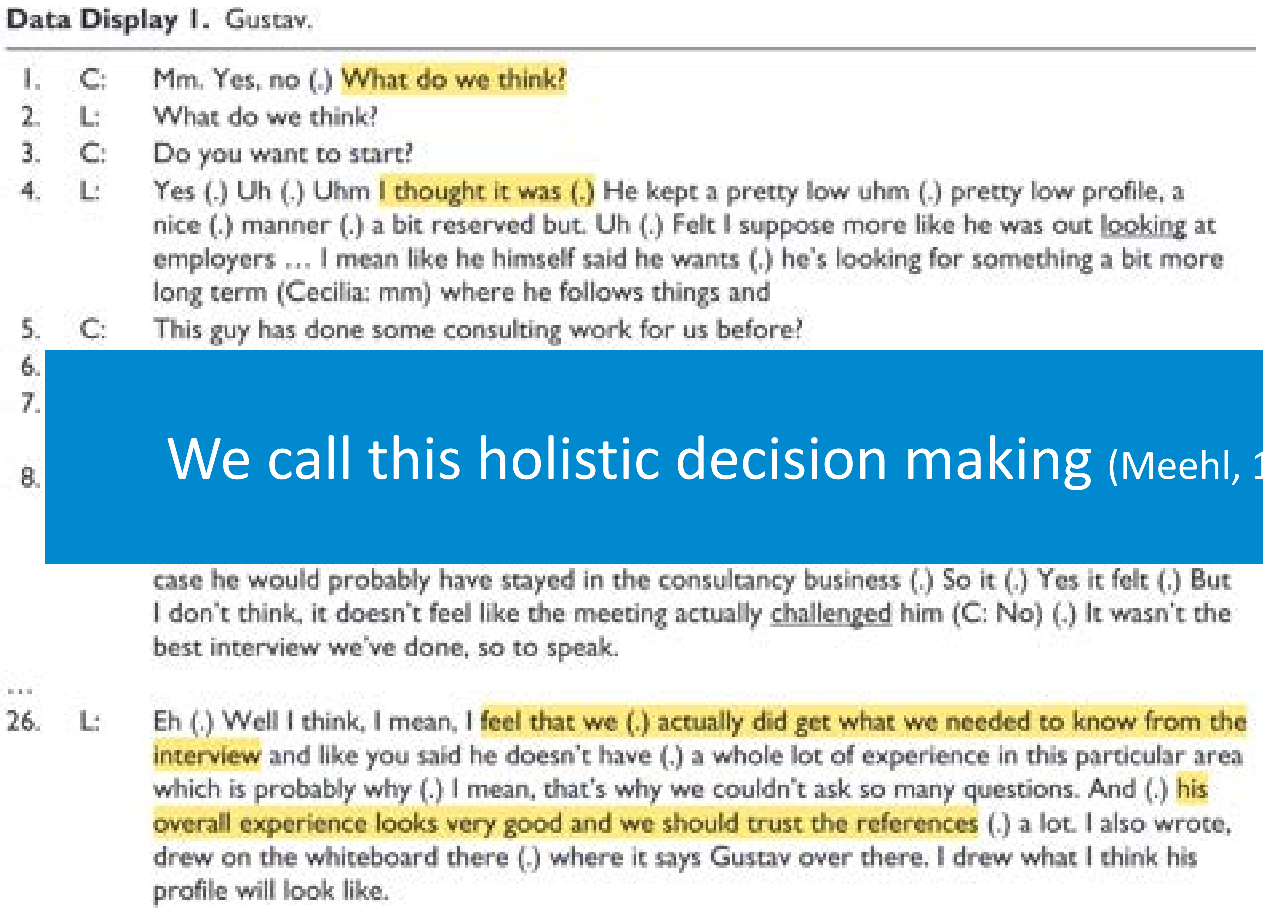
Imagine this Situation…

Hm… What do I think about this candidate?



Interpersonal Communication in Selection

- We constantly explain ourselves (Bolander & Sandberg, 2013)



Data Display 1. Gustav.

1. C: Mm. Yes, no (.) What do we think?
2. L: What do we think?
3. C: Do you want to start?
4. L: Yes (.) Uh (.) Uhm I thought it was (.) He kept a pretty low uhm (.) pretty low profile, a nice (.) manner (.) a bit reserved but. Uh (.) Felt I suppose more like he was out looking at employers … I mean like he himself said he wants (.) he's looking for something a bit more long term (Cecilia: mm) where he follows things and
5. C: This guy has done some consulting work for us before?
6.
7.
8.

case he would probably have stayed in the consultancy business (.) So it (.) Yes it felt (.) But I don't think, it doesn't feel like the meeting actually challenged him (C: No) (.) It wasn't the best interview we've done, so to speak.

…

26. L: Eh (.) Well I think, I mean, I feel that we (.) actually did get what we needed to know from the interview and like you said he doesn't have (.) a whole lot of experience in this particular area which is probably why (.) I mean, that's why we couldn't ask so many questions. And (.) his overall experience looks very good and we should trust the references (.) a lot. I also wrote, drew on the whiteboard there (.) where it says Gustav over there. I drew what I think his profile will look like.

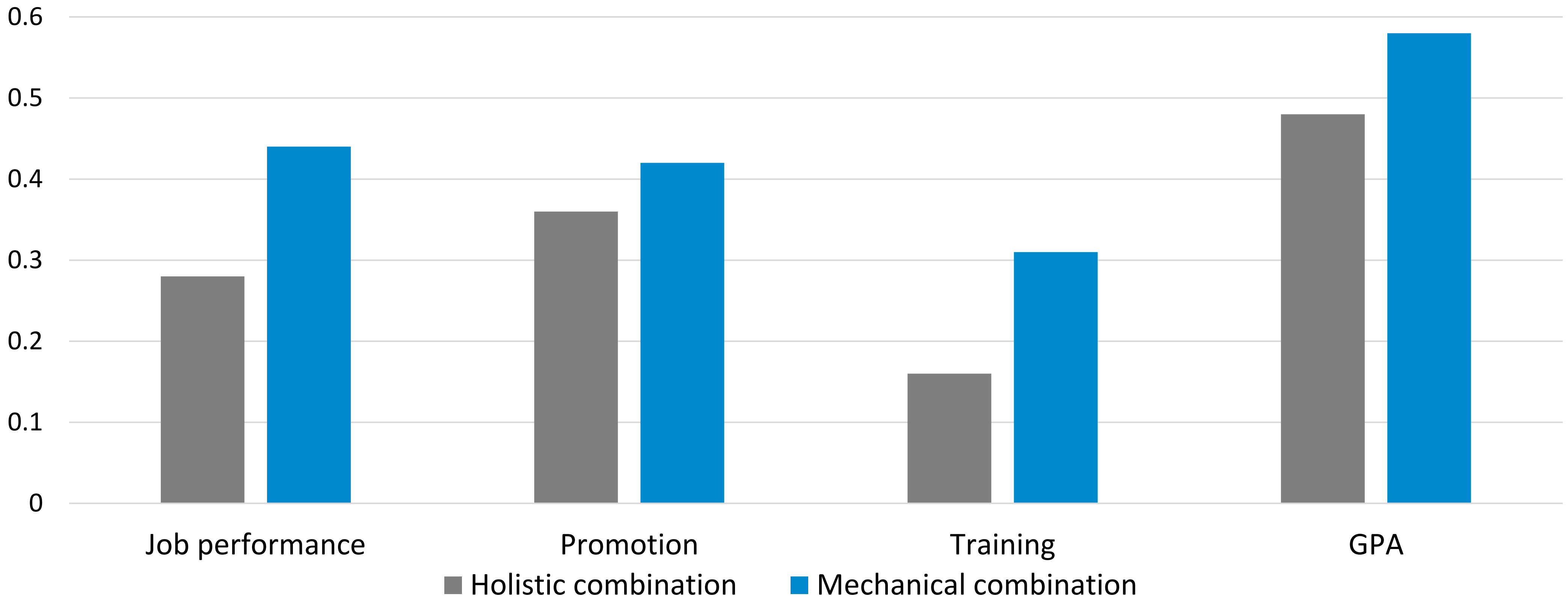
We call this holistic decision making (Meehl, 1954)

Mechanical Decision Making

- Use an algorithm or rule to combine information (Meehl, 1954)
  - Fit rating = GMA*1 + Conscientiousness*1 + Interview rating*1
  - Fit rating = GMA*0.7 + Conscientiousness*0.2 + Interview rating*0.1
  - Hire if GMA >= 100 and Interview rating everything else than the worst

(Simple) Algorithms Beat Expert Judgment

Predictive Validity per Combination Method (Kuncel et al., 2013)



Algorithms as Advisors

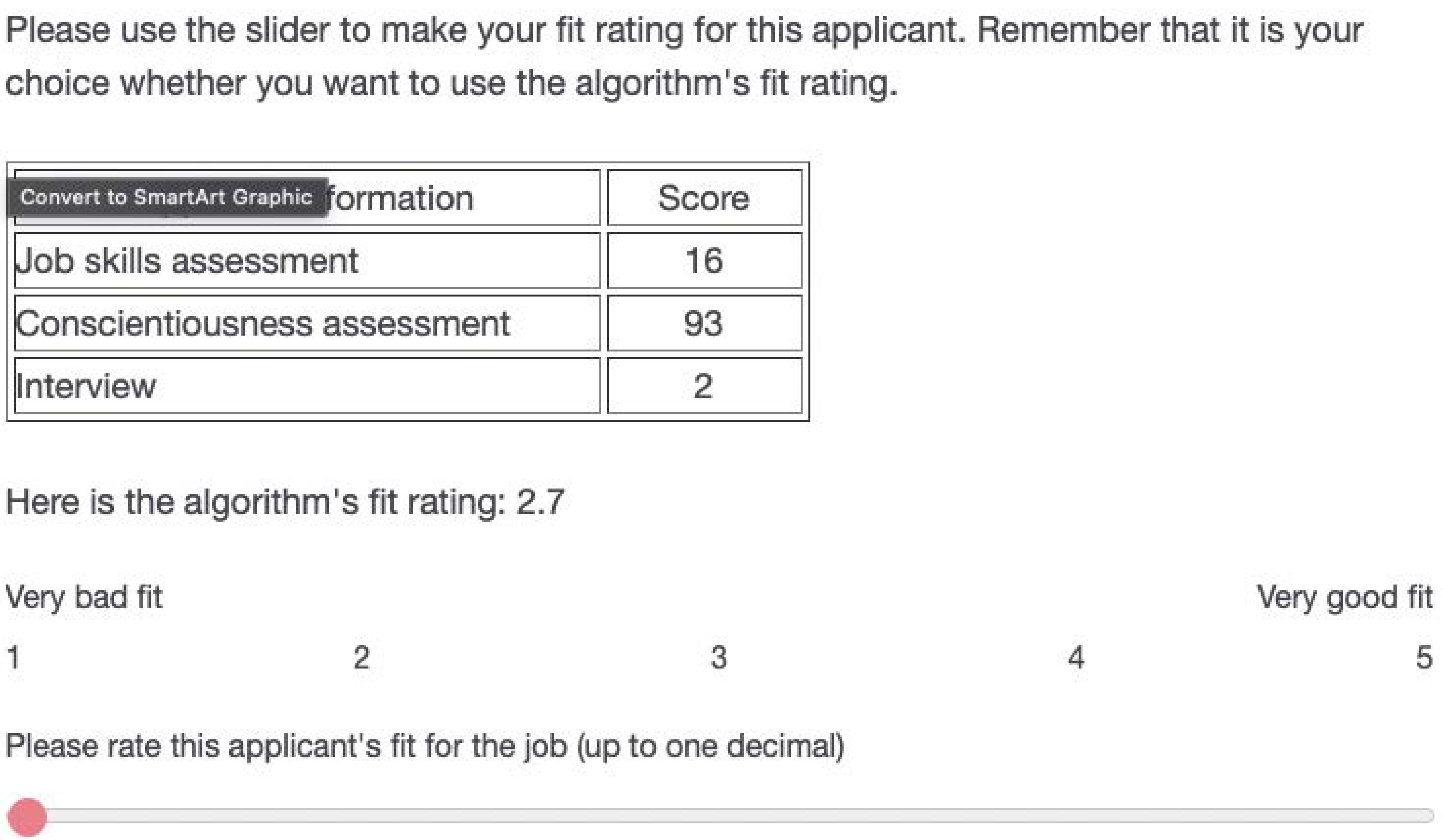
- When algorithms are used at all, predictions serve as mere advice
- Considering algorithmic advice…

1. Increases validity compared to **pure holistic prediction** (Dietvorst et al., 2018; Neumann et al., 2022; 2023)
2. Decreases validity compared to **strict algorithm use**

Algorithms as Advisors

- RQ: How can we increase the consistent use of algorithmic advice?
  - Have an algorithm “explain itself” –> tell (data-based) stories
  - Stimulate decision-makers’ sense making of algorithmic predictions

Participants and Prediction Task



Please use the slider to make your fit rating for this applicant. Remember that it is your choice whether you want to use the algorithm's fit rating.

formation	Score
Job skills assessment	16
Conscientiousness assessment	93
Interview	2

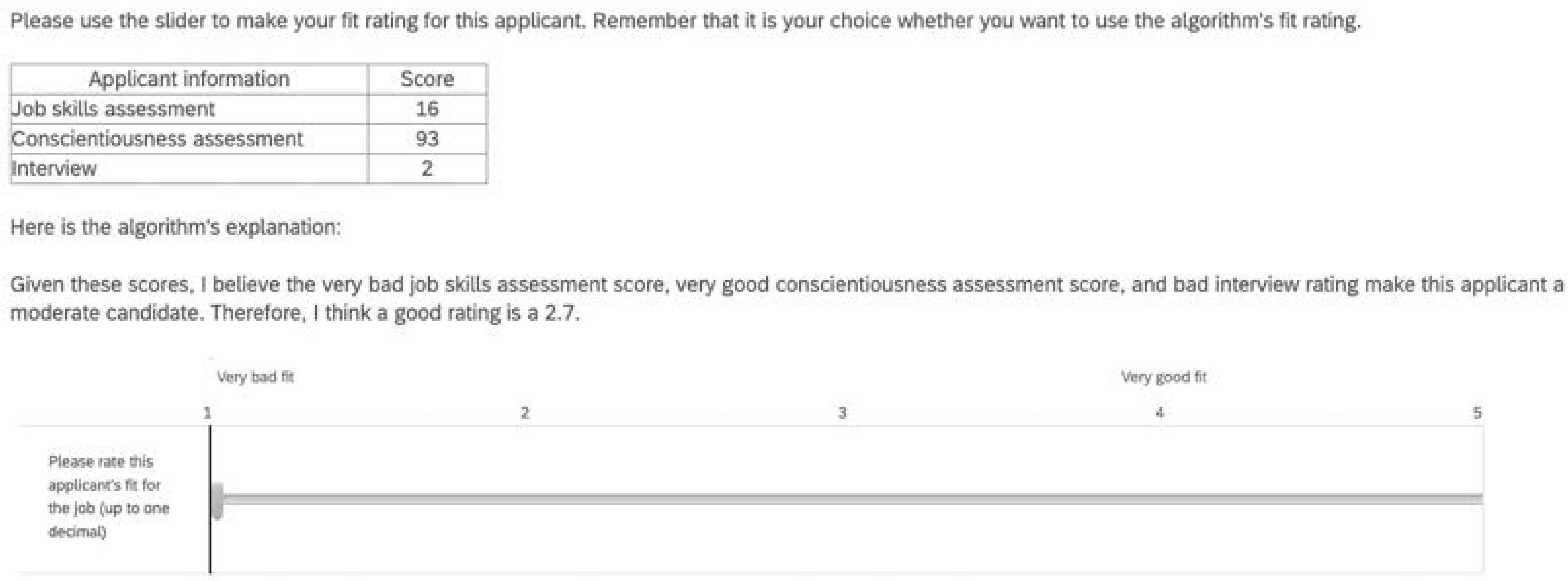
Here is the algorithm's fit rating: 2.7

Very bad fit Very good fit

1 2 3 4 5

Please rate this applicant's fit for the job (up to one decimal)

Design



Please use the slider to make your fit rating for this applicant. Remember that it is your choice whether you want to use the algorithm's fit rating.

Applicant information	Score
Job skills assessment	16
Conscientiousness assessment	93
Interview	2

Here is the algorithm's explanation:

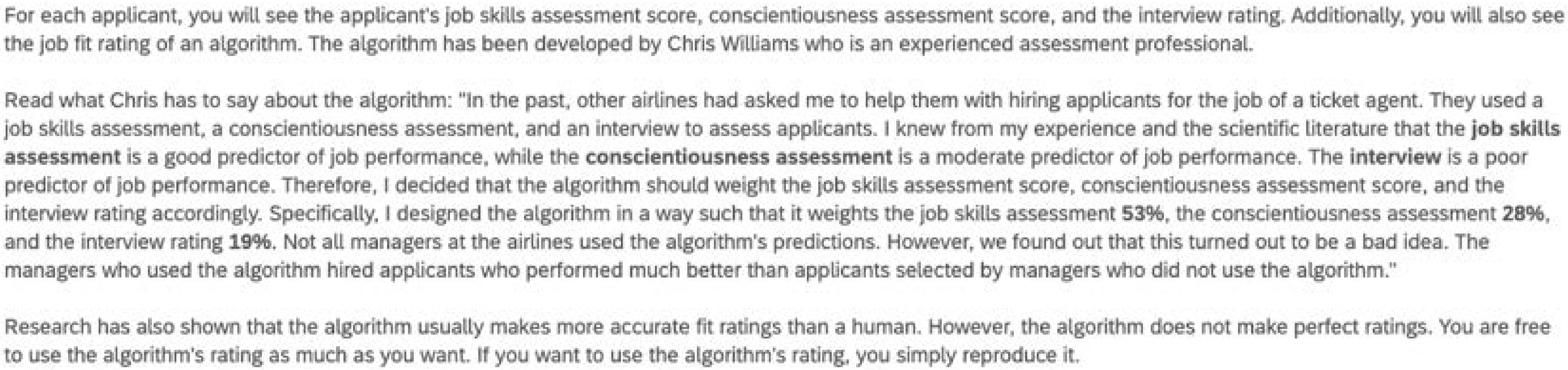
Given these scores, I believe the very bad job skills assessment score, very good conscientiousness assessment score, and bad interview rating make this applicant a moderate candidate. Therefore, I think a good rating is a 2.7.

Please rate this applicant's fit for the job (up to one decimal)

Algorithm’s Explanation

1. Set approximately equal score bands on predictor scores
2. Translate scores into words (interview = 2 -> “bad”)
3. Randomly vary the intro, verb, and end of a sentence
   - Intro = “In this case,”, “Given these results,”, “Looking at these numbers,”, “Based on this profile,”
   - Verb = “think”, “would say”, “believe”
   - End = “my rating for this applicant would be a”, “I would give this applicant a rating of”, “my rating in this case would be a”, “I think a good rating is a”
4. Knit everything together

Design

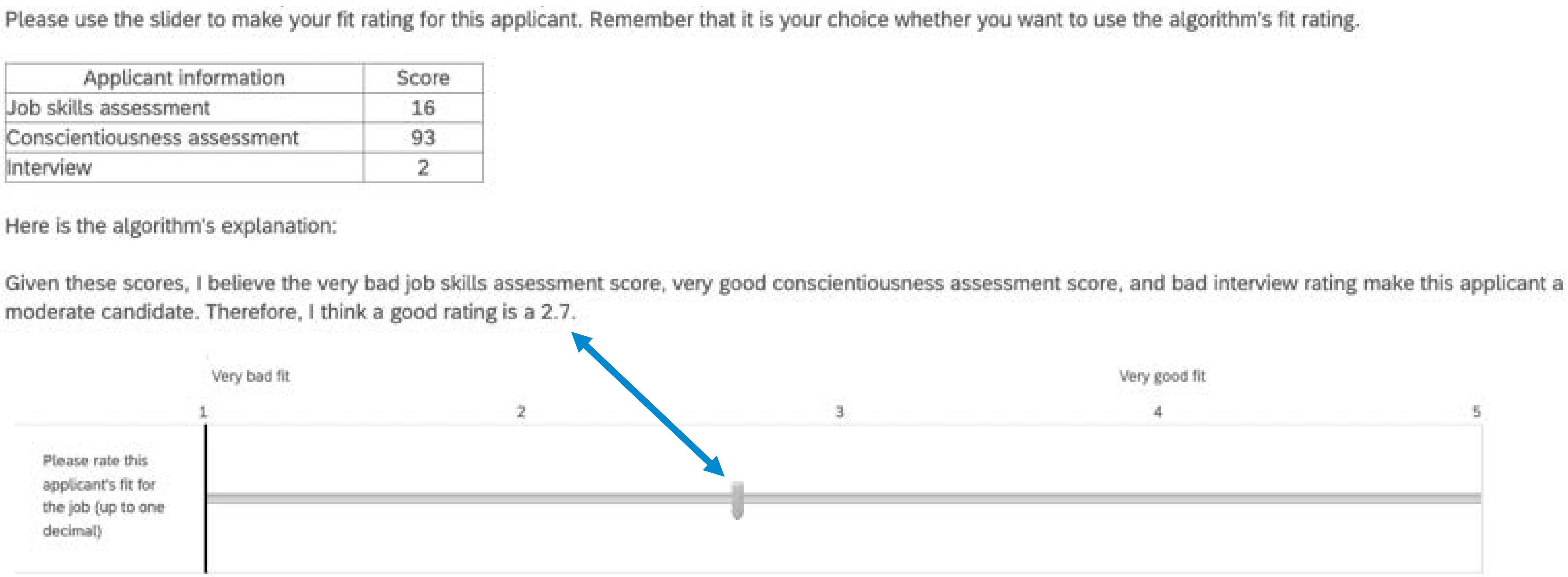


For each applicant, you will see the applicant's job skills assessment score, conscientiousness assessment score, and the interview rating. Additionally, you will also see the job fit rating of an algorithm. The algorithm has been developed by Chris Williams who is an experienced assessment professional.

Read what Chris has to say about the algorithm: "In the past, other airlines had asked me to help them with hiring applicants for the job of a ticket agent. They used a job skills assessment, a conscientiousness assessment, and an interview to assess applicants. I knew from my experience and the scientific literature that the **job skills assessment** is a good predictor of job performance, while the **conscientiousness assessment** is a moderate predictor of job performance. The **interview** is a poor predictor of job performance. Therefore, I decided that the algorithm should weight the job skills assessment score, conscientiousness assessment score, and the interview rating accordingly. Specifically, I designed the algorithm in a way such that it weights the job skills assessment **53%**, the conscientiousness assessment **28%**, and the interview rating **19%**. Not all managers at the airlines used the algorithm's predictions. However, we found out that this turned out to be a bad idea. The managers who used the algorithm hired applicants who performed much better than applicants selected by managers who did not use the algorithm."

Research has also shown that the algorithm usually makes more accurate fit ratings than a human. However, the algorithm does not make perfect ratings. You are free to use the algorithm's rating as much as you want. If you want to use the algorithm's rating, you simply reproduce it.

Design



Please use the slider to make your fit rating for this applicant. Remember that it is your choice whether you want to use the algorithm's fit rating.

Applicant information	Score
Job skills assessment	16
Conscientiousness assessment	93
Interview	2

Here is the algorithm's explanation:

Given these scores, I believe the very bad job skills assessment score, very good conscientiousness assessment score, and bad interview rating make this applicant a moderate candidate. Therefore, I think a good rating is a 2.7.

Please rate this applicant's fit for the job (up to one decimal)

Measures

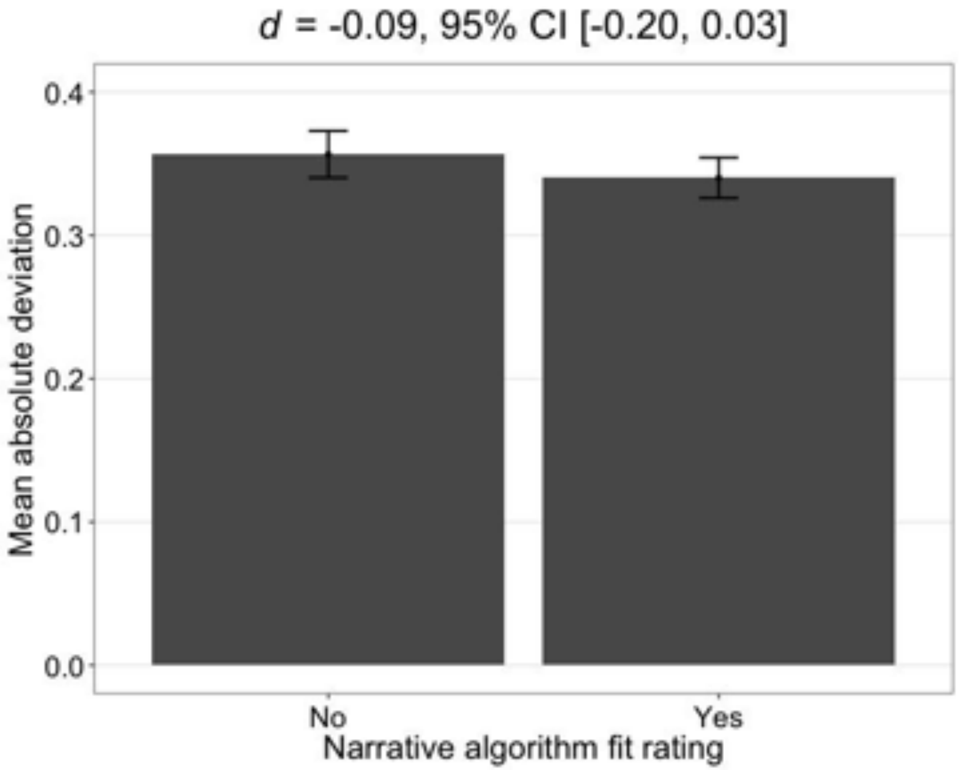
- Algorithm use: $Mean\ absolute\ deviation\ = \frac{\sum_i^{40}|Pi-Ai|}{40}.$
- Judgment consistency: Fit rating ~ the three predictors
- Validity: *r* between the 40 fit ratings and performance ratings
- Attitudinal measures
  - Trust (”I have trust in the algorithm’s fit ratings”)
  - Anthropomorphism (”I felt like I was interacting with a human when making fit ratings”)
  - Use intentions (”I would choose to use the algorithm to make future hiring decisions”)

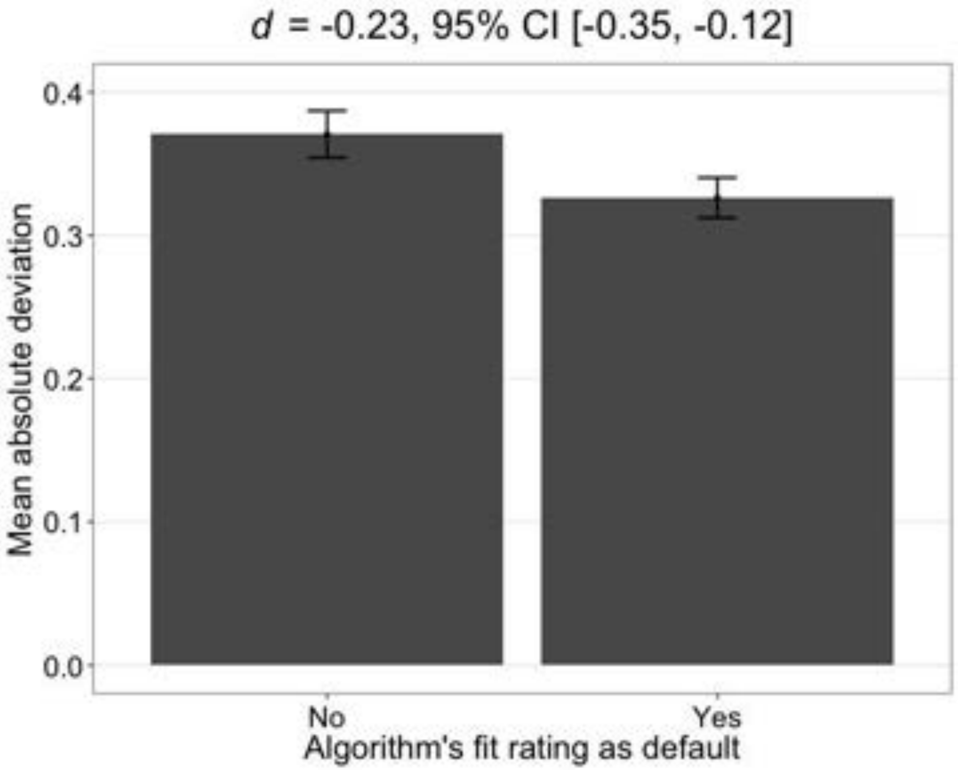
Results

Mean optimal model validity	Mean algorithm validity	Mean participant validity
.42	.34	.32

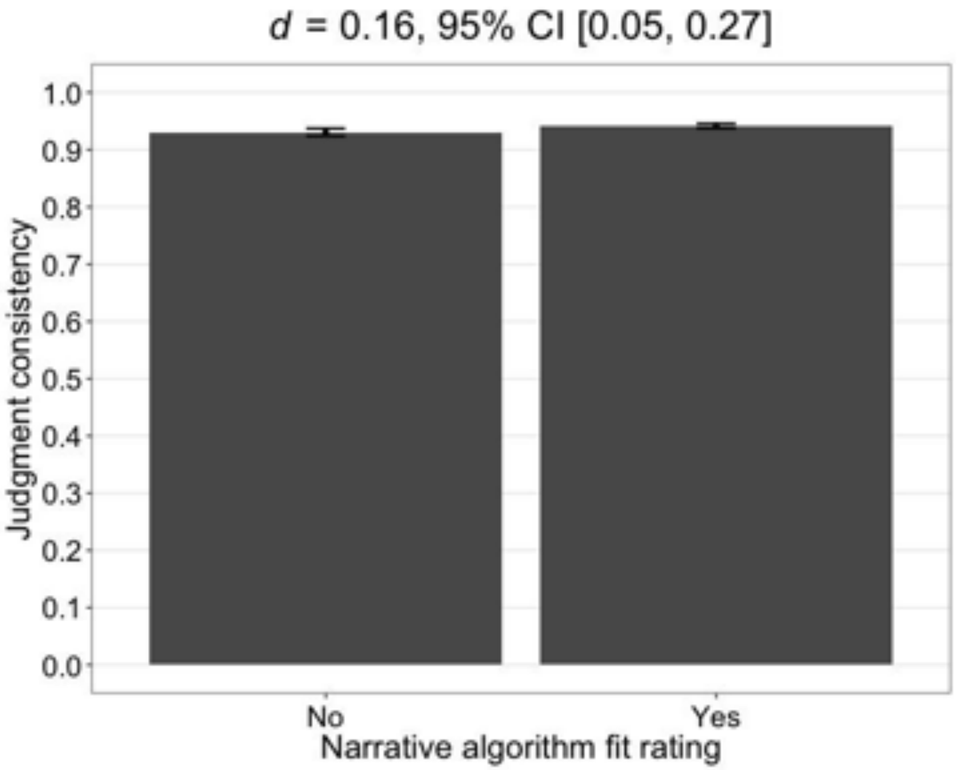
- Participants’ validity was slightly lower than algorithm’s validity
  - This left little room for our interventions to improve participants’ validity

Mean Absolute Deviation

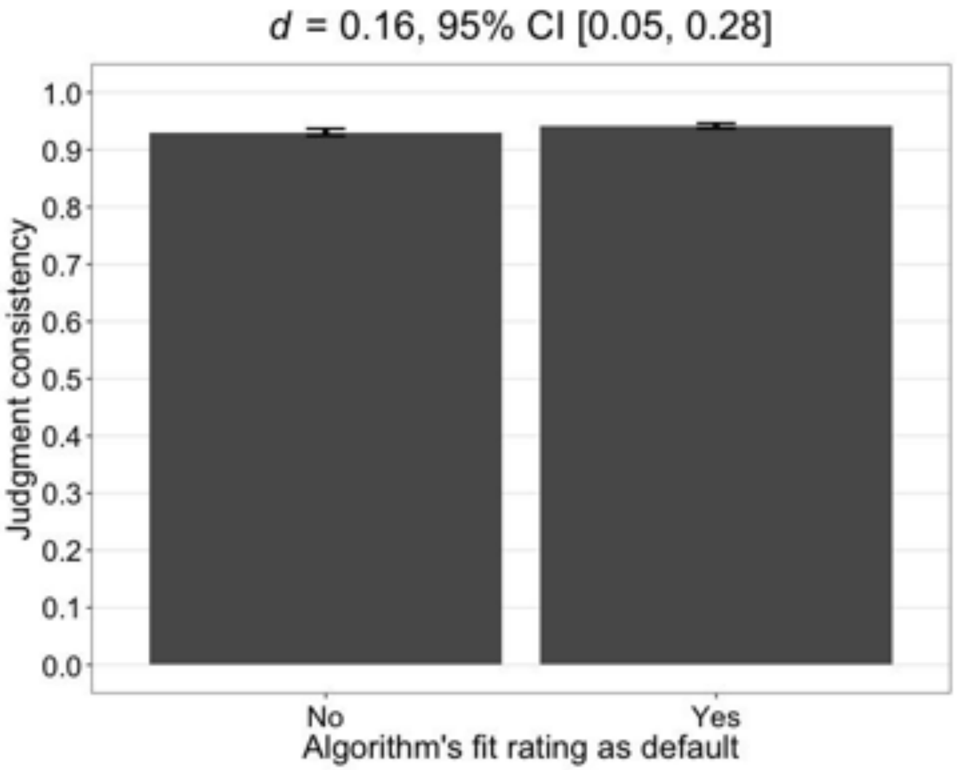




Judgment Consistency

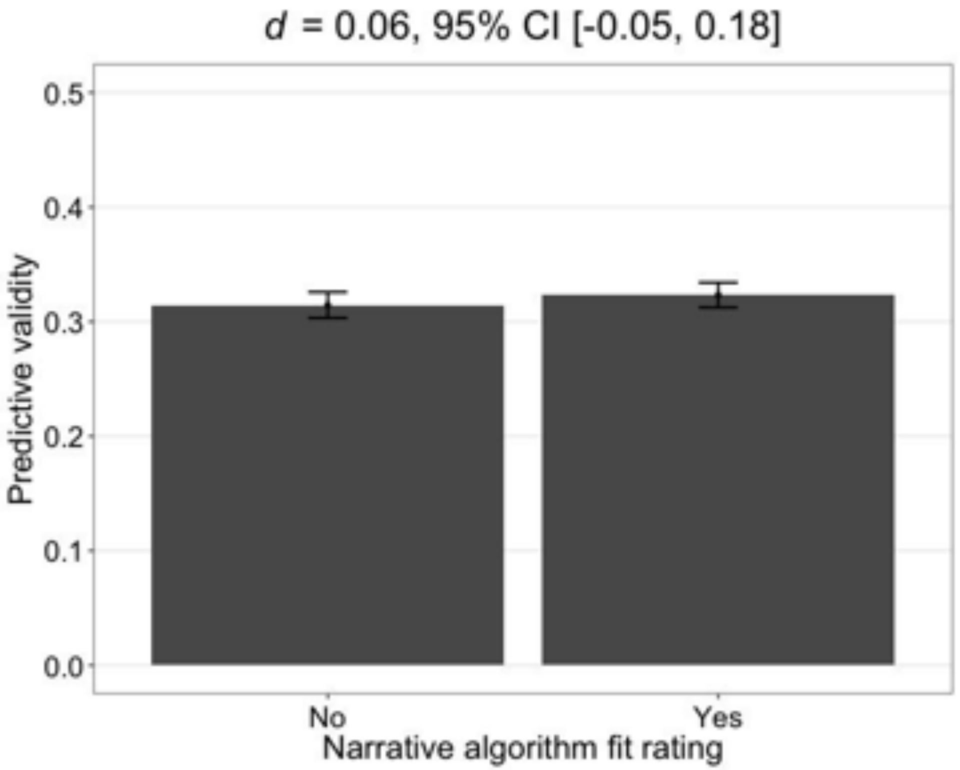


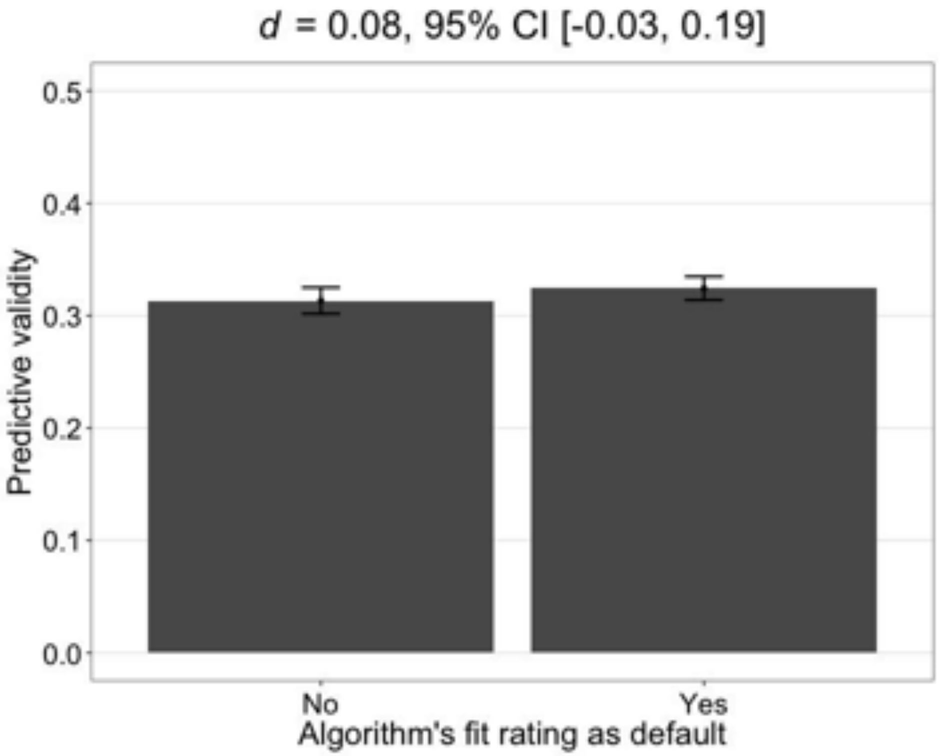
d = 0.16, 95% CI [0.05, 0.27]



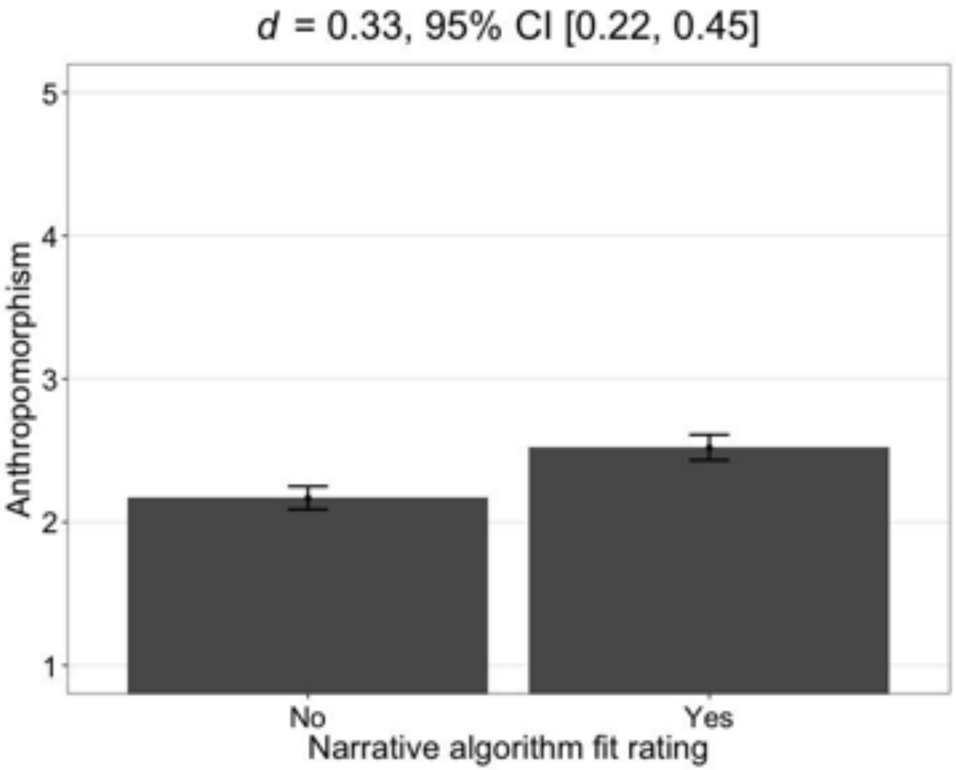
d = 0.16, 95% CI [0.05, 0.28]

Predictive Validity

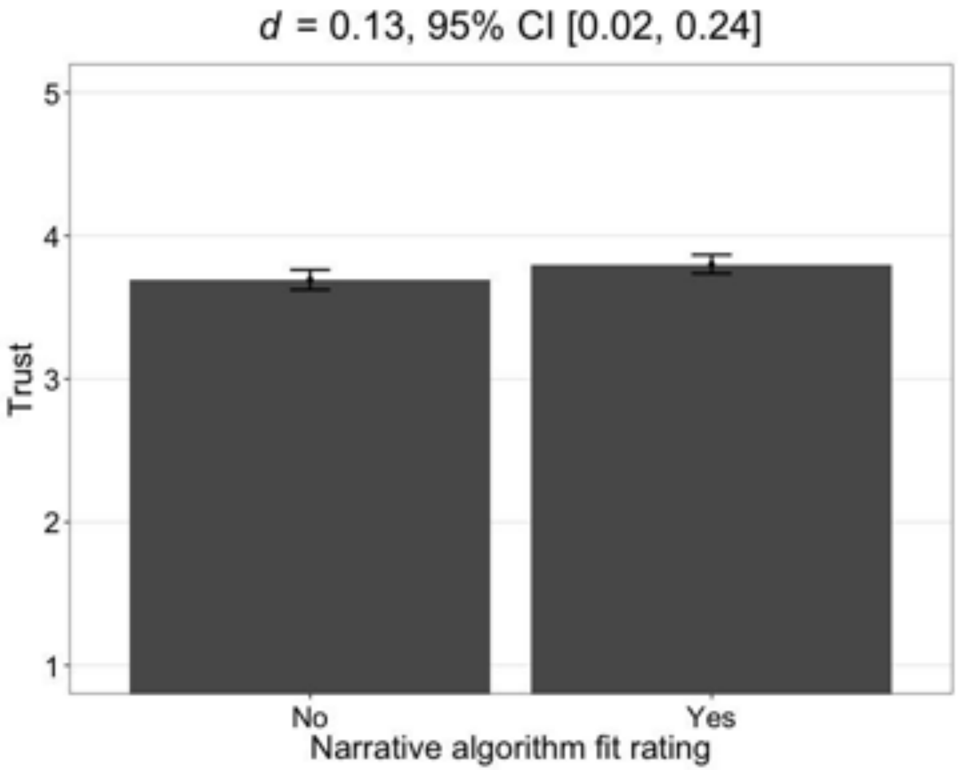




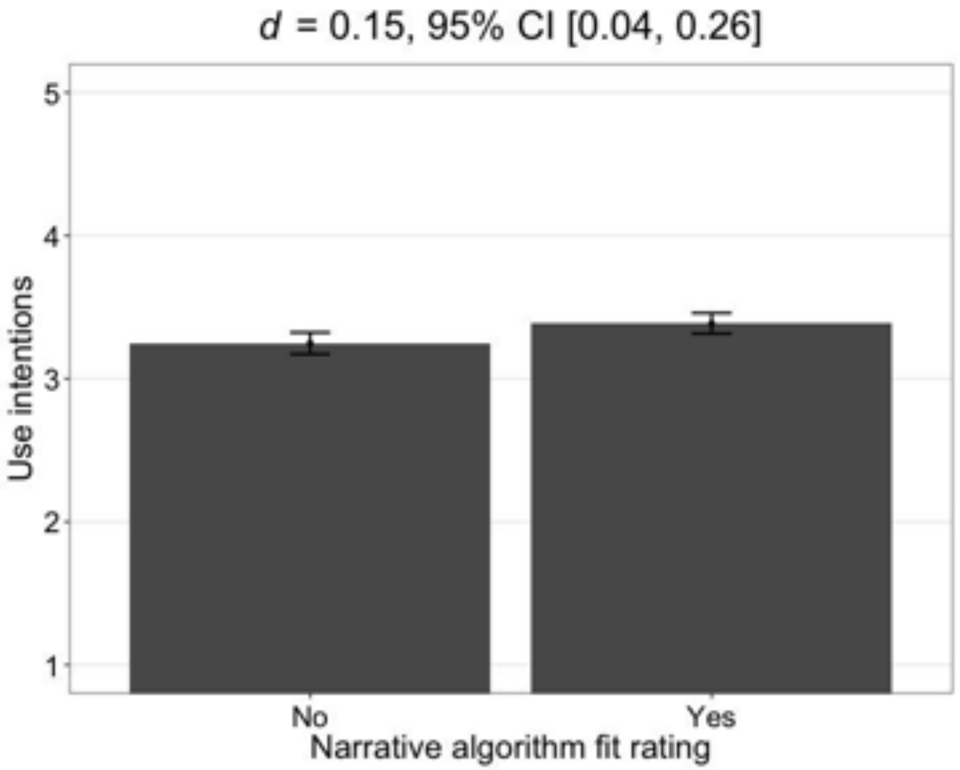
Anthropomorphism



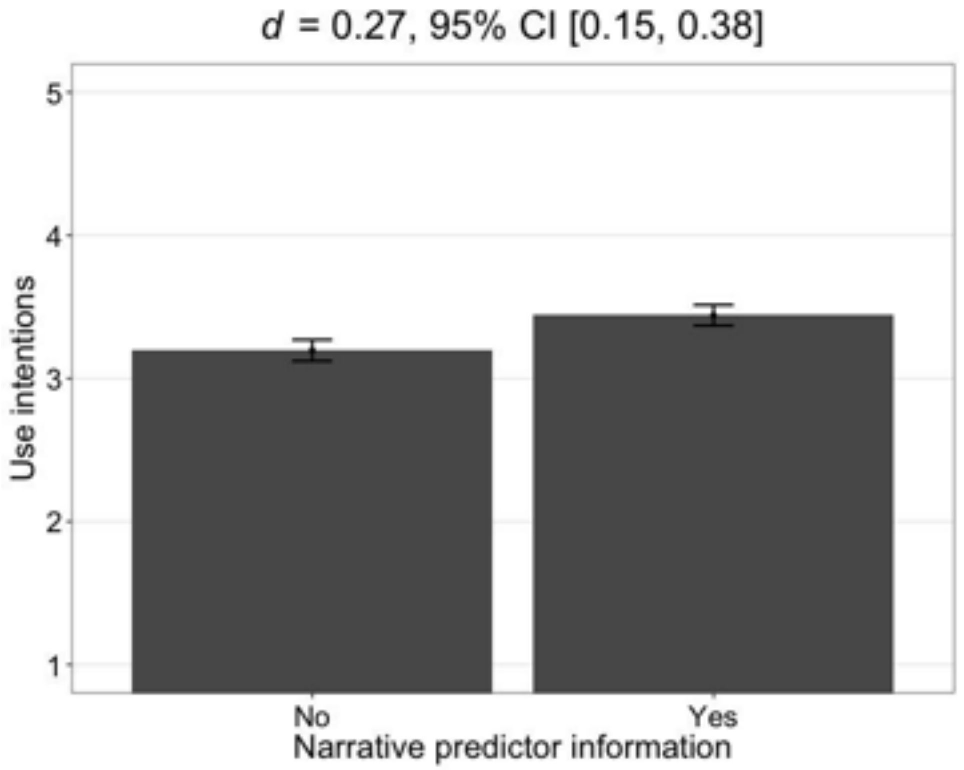
Trust


d = 0.13, 95% CI [0.02, 0.24]
Trust
5
4
3
2
1
No
Yes
Narrative algorithm fit rating

Use Intentions



d = 0.15, 95% CI [0.04, 0.26]



d = 0.27, 95% CI [0.15, 0.38]

Discussion

- We used a low-key explanation: Short, descriptive, text, no avatar
- Our algorithm did not think nor type (think of ChatGPT)
  - No interaction/two-way communication -> less sense making?
- No qualitative predictor information
  - More reason to deviate
  - Richer stories
- How can decision makers make sense of algorithmic predictions?

Thank you for your attention!

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

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

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

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

Neumann, M., Niessen, A. S. M., Linde, M., Tendeiro, J. N., & Meijer, R. R. (2023). “Adding an egg” in algorithmic decision making: Improving stakeholder and user perceptions, and predictive validity by enhancing autonomy. *European Journal of Work and Organizational Psychology*, Advance online publication. https://doi.org/10.1080/1359432X.2023.2260540

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