Text S1

Response Validity Analyses

With its extensive length of 300 items, the Big Five Structure Inventory (BFSI; Arendasy, 2009) has an increased risk of triggering careless or insufficient effort (C/IE) responding (Curran, 2016; Ward & Meade, 2023). Hence, we conducted a medium-level analysis of response validity, as recommended by Ward and Meade (2023). Please note that we performed these analyses post-hoc (i.e., after our machine learning benchmarks) to estimate the impact of careless responding on our predictions and not to remove participants beforehand. To detect different forms of careless responding, we combined a) multivariate outlier analysis via Mahalanobis Distance, invariance analysis via b) the long-string index, and c) intraindividual response variability (IRV), and d) consistency analysis via the even-odd index (Johnson, 2005; Meade & Craig, 2012). All analyses were conducted with the careless package in R (Yentes & Wilhelm, 2021), and the respective code is available in our project repository under https://osf.io/x7dar/. Mahalanobis Distance revealed no multivariate outliers, indicating that none of our participants exhibited aberrant responses across all items. The long-string analysis detected 32 cases with over 15 identical responses in consecutive items, two of which lasted for over 40 items. Similarly, IRV identified seven participants whose intra-individual standard deviation across items was more than two standard deviations below the sample's mean. While these two indices seem to flag some participants as careless, invariability should not be over-interpreted in the context of the BFSI, which contains 60 items assessing the same Big Five dimension and consists only of adjectives with the same directionality and intensity (Dunn et al., 2018). Finally, the even-odd consistency was critically low (i.e., below the recommended cutoff of .30) for only one participant, indicating a lack of consistent responses within the BFSI's sub-scales. In sum, we cannot rule out that our analyses also included instances of invalid data produced through careless responding.

However, we refrained from removing these participants for a lack of appropriate and unambiguous evidence and discuss the limited response validity instead (Curran, 2016).

References

Arendasy, M. (2009). *BFSI: Big-Five Struktur-Inventar (Test & Manual).* SCHUHFRIED GmbH.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology, 33*, 105-121. https://doi.org/10.1007/s10869-016-9479-0

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 103-129. https://doi.org/10.1016/j.jrp.2004.09.009

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology, 74*, 577-596. https://doi.org/10.1146/annurev-psych-040422-045007

Yentes, R. D., & Wilhelm, F. (2021). *careless: Procedures for computing indices of careless Responding* (Version 1.2.1) [R package].