Advancing Perspectivist Ground Truthing with Social Science

Table of contents

## Poor reproducibility/replicability in ML research

Hullman et al. (2022) compare claims that ML is facing a reproducibility crisis to the crisis in psychology. Among the issues they note relate to reference datasets.

They note that researchers often re-use datasets as they publish on standardized benchmarks, and because they are cost prohibitive to collect. It may also be the case that ML researchers broadly prefer to work on building and evaluating performance, rather than executing ground-truthing projects Sambasivan et al. (2021).

* use of non-representative samples
* measurement bias
* underspecification of the population captured in data
* transformation of data to optimize for ‘accuracy’

With caveats, they note that there are lessons to be learned from social science, and their replication crisis.

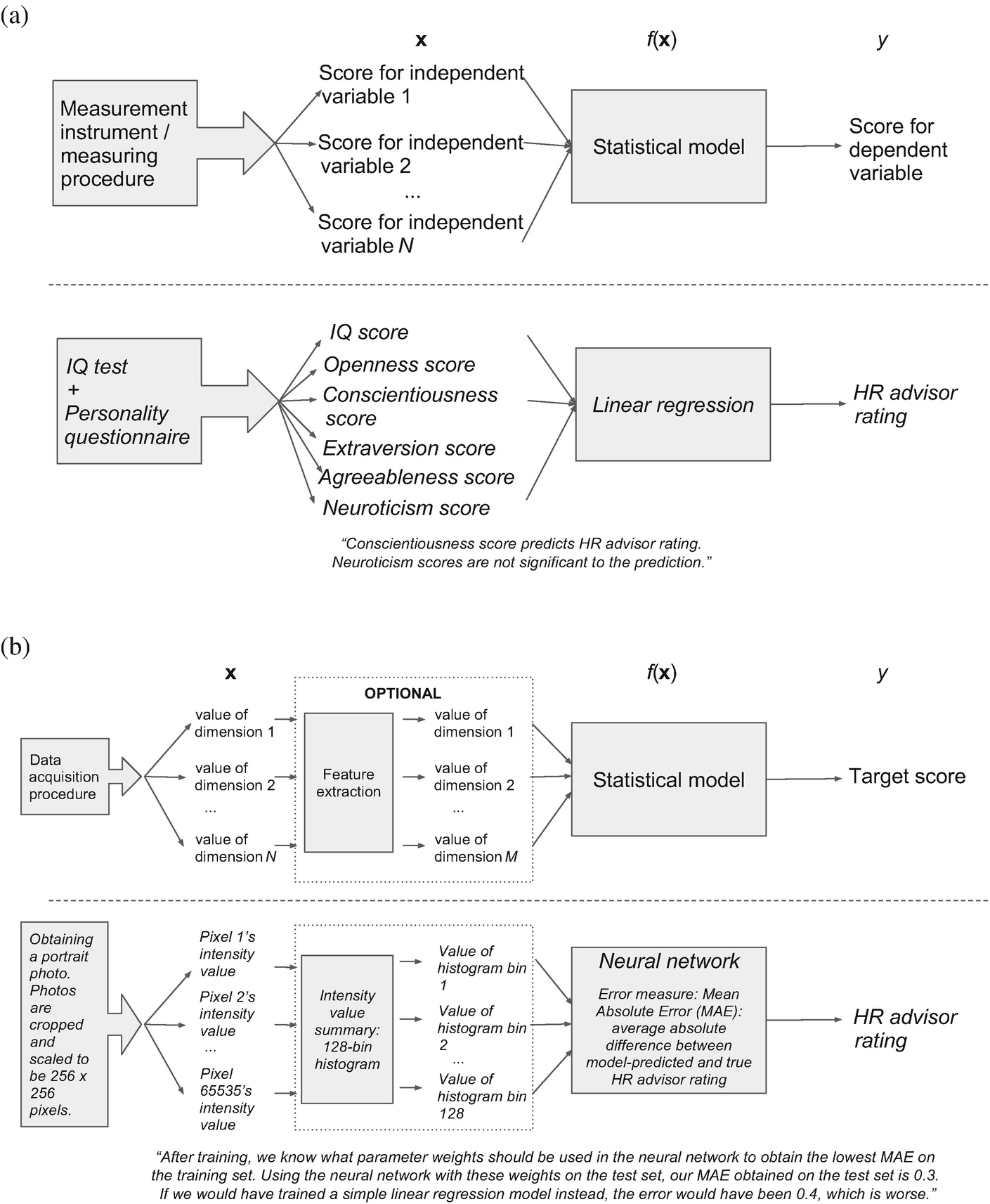
### Social and computational sciences have different focci

social sciences:

* interpretable meaning of x and y
* design is informed by theory

computational sciences:

* learning procedure f(x)

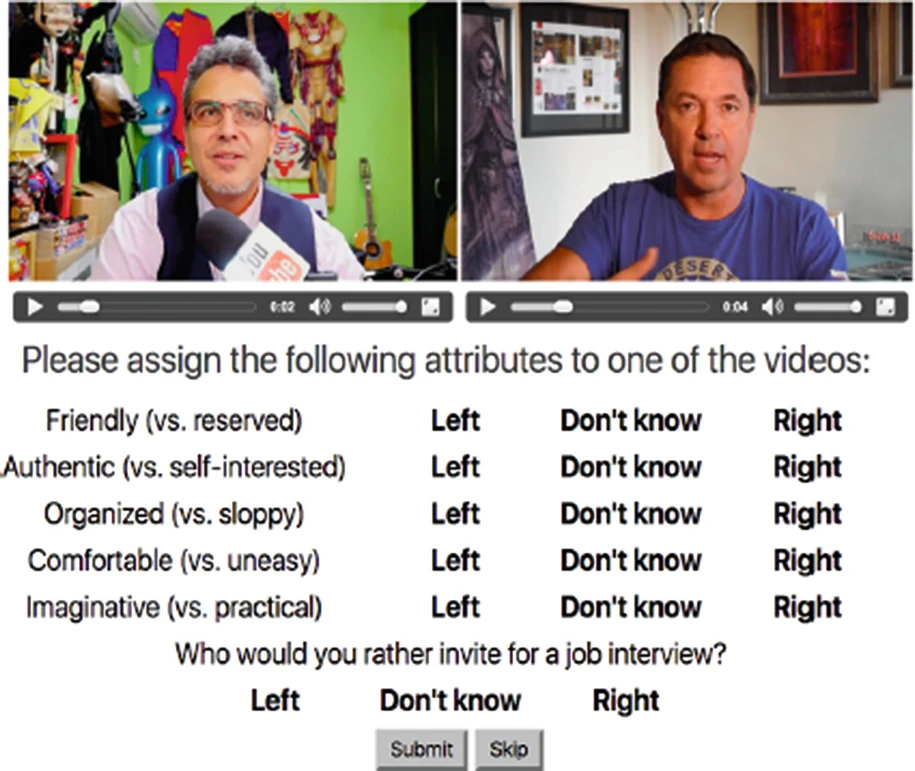
 Fig. 1: Liem et al. (2018)

thus, positive yet unrealized interdisciplinary potentia, but also strong negative potential when disciplines ignore fundamental scientific issues in their research pipelines i.e. the “worst of both worlds” Hullman et al. (2022)

#### A tale of two studies: the case of personnel selection

Ponce-López et al. (2016)

* efficient way to gather media and annotation data
* BUT no validation of instrument, or ecologically valid media data

 Figure 2: Ponce-López et al. (2016)

Compared to Koutsoumpis et al. (2024):

* ecological validity: media data was mock asynchronous video interviews
* ecological validity: interview questions designed to activate personality facets
* personality instruments: validated HEXACO scale
* perspectives: self & observer ratings

### ML uses human annotations very often

Geiger et al. (2021)

* 200 randomly sampled ML papers from 3 domains:
  + Social Sciences & Humanities
  + Life & Biomedical Sciences
  + Physical & Environmental Sciences
* Out of 141 classification tasks, 103 (73.05%) used human labels
* Out of 103 human labels, 58 (56.31%) used only external labels

i.e. ML re-uses external labels, and inadequately reports ‘ground truth’

without details of ground truth, we cannot know what the resulting model represents [perhaps cat image parable here?]

### Human annotations aren’t always accurate

Griffin & Brenner (2004) review errors and biases in human judgements[[1]](#footnote-29)

* over/under prediction: confidence score is higher/lower than accuracy
* over/under extremity: confidence is more extreme at ends

also reviews possible reasons:

* optimistic overconfidence
* confirmation bias
* case-based judgment
* ecological probability
* error model (psychometric model)

### ML often ignores perspectives of annotators

[crowd truth; perspectivist approach here]

#### Issues with sampling

Solutions to sampling problems can come from sampling theory: Groves et al. (2009)

considerations:

* sampling frame: the elements in from populations that you have access to
* ineligible units: elements in the sampling frame that are not your target
* undercoverage: elements from target population that are not in the frame

solutions:

* stratified sampling

#### Issues with instruments

Beck et al. (2022)

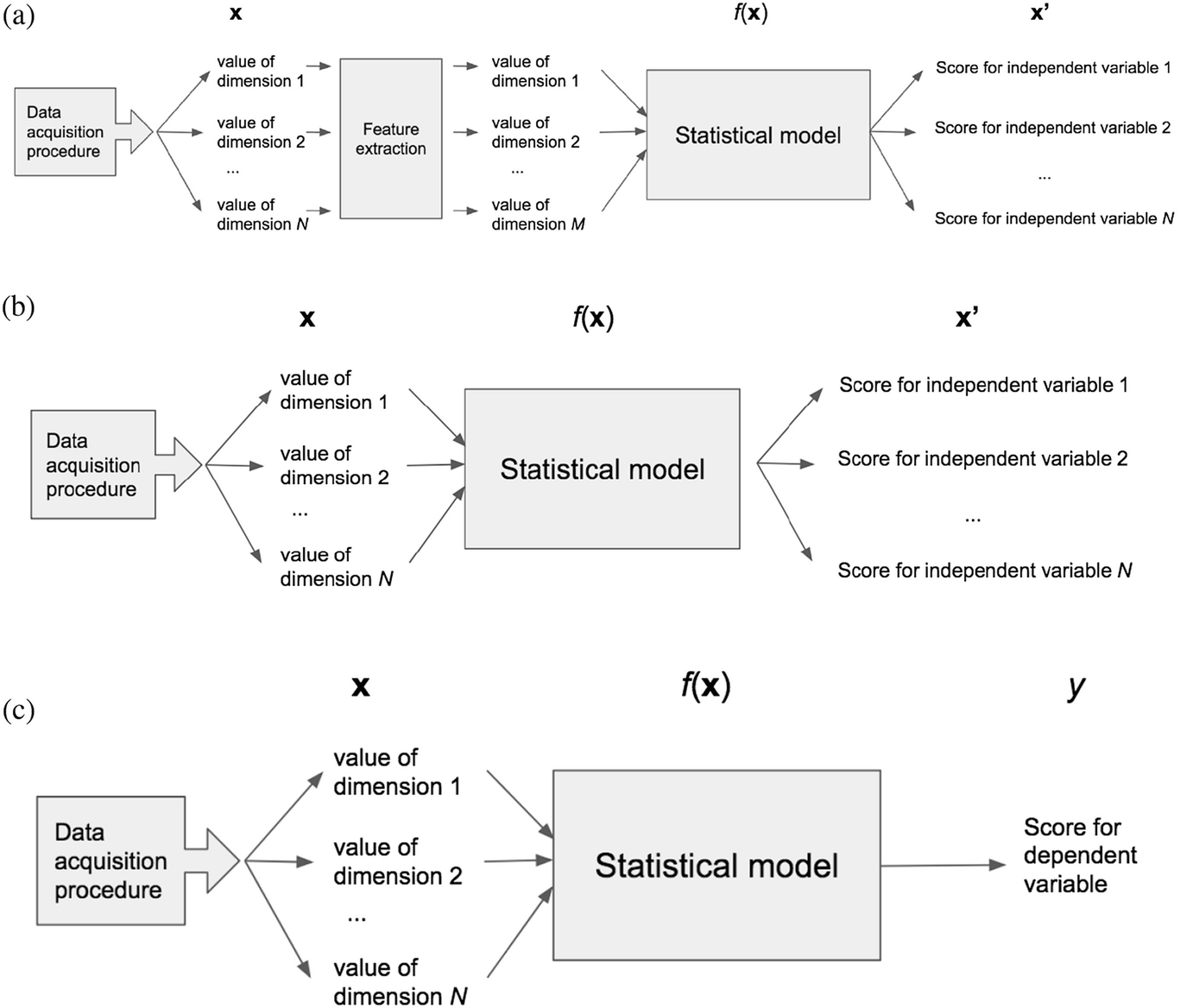
* annotation collection requires design thinking
  + Task Structure: specific wording and response options, including debates over the inclusion of “I don’t Know” option
  + Order Effects: specific judgements are affected by previous perceptions
  + Annotator Effects: backgrounds, opinions, experiences of respondents affect responses

[another ref that describes the target as a latent variable]

## Present Work

We incorporate these considerations in the design of our study, and attempt to further the field in the following ways:

* we attempt representative sampling of both media and respondents
* we aim to estimate 10-dimensional psychological construct
* we select media that is ambiguous (i.e. that will result in subjectivity in the ratings) as well as media that we expect not to be ambiguous for comparison
* we estimate a-priori the number of ratings necessary rather than assuming
* we take into account perspectives

case study of this thesis works towards path (b) in Liem et al. (2018) shown in: 

## References

Beck, J., Eckman, S., Chew, R., & Kreuter, F. (2022). Improving labeling through social science insights: Results and research agenda. *International Conference on Human-Computer Interaction*, 245–261.

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). " garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *arXiv Preprint arXiv:2107.02278*.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. *Blackwell Handbook of Judgment and Decision Making*, *199*, 158–177.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (Vol. 561). John Wiley & Sons.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348.

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., Breda, W. van, Zhang, T., & Vries, R. E. de. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. *Computers in Human Behavior*, *154*, 108128.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 197–253.

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J., & Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. *Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, 400–418.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

1. Griffin & Brenner (2004) note that much of this work was about people guessing knowledge from an almanac, and then guessing how accurate they were [↑](#footnote-ref-29)