



Comment on "Responsible Research Assessment: Implementing DORA for hiring and promotion in psychology".

Victor Auger¹ and Nele Claes¹

¹Université Clermont Auvergne, CNRS, LAPSCO, Clermont-Ferrand, France

In target papers, Schönbrodt et al. (2022) and Gärtner et al. (2022) proposed to broaden the range of the considered research contributions, namely (i) bringing strong empirical evidence, (ii) building open databases, (iii) building and maintaining packages, where each dimension being scored independently in marking scheme. Using simulations, we show that the current proposal places a significant weight on software development, potentially at the expense of other academic activities – a weight that should be explicit to committees before they make use of the proposed marking scheme. Following Gärtner et al. (2022) recommendations, we promote the use of flexible weights which more closely match an institution's specific needs by the weighting of the relevant dimensions. We propose a Shinyapp that implement the marking scheme with adaptative weights to both help the hiring committee define and foresee the consequences of weights' choices and increase the transparency and understandability of the procedure.

Keywords: Research Assessment, Simulation-based Approach, Open Science

The academic hiring process is a cornerstone of the scientific progress. Throughout two papers, Gärtner et al. (2022) and Schönbrodt et al. (2022) criticize current indicators used in the hiring process and propose a new way to assess candidates following the San Francisco Declaration on Research Assessment (DORA) considerations for Open Science practices. We agree with the need to reform the evaluative procedure and welcome the proposed implementation to lead the field toward a robust psychological science by adjusting how researchers are evaluated, and then hired or not. In the proposed implementation, the authors considered three main contributions that promote research quality, namely (i) bringing strong empirical evidence (papers contribution), (ii) building open databases (data contribution), (iii) building and maintaining packages (software contribution). Through a well-detailed operationalization, candidates earn points based on their contributions in each dimension. Then, candidates that meet a minimum threshold are considered for a more detailed evaluation. While the first phase is described as "negative selection" (i.e., to reject candidates below a minimum required; see Figure 2 on p. 6; Schönbrodt et al., 2022), the ratio between the increasing number of candidates and the hiring committee's member time constraints will mechanically lead to an increase of the threshold value. This further underscores the significance of the first evaluation stage during which a large proportion of candidates will be rejected. Given the cen-

trality of this first step, we conducted a simulation study to assess how the different dimensions of the implementation would rank against each other.

Simulation-based assessment of the implementation

We followed the implementation proposed by Gärtner et al. (2022) to assess the importance of paper contributions, data contributions, and software contributions. Each dimension was respectively scored on a scale 1-12, 1-5, and 1-24 points. Candidates were assigned a profile of low, mid, or high-scorer for each dimension. Mean scores and standard deviations were calculated based on the maximum score for each dimension, resulting in 27 (33) different candidate profiles (see Supplementary Materials). We then simulated, ranked, and plotted all plausible candidates and examined how the different profiles would fare against each other.¹ Our descriptive analysis indicates that the score for software contribution has a strong influence on the ranking, almost independently from the two other contributions. In other words, the candidates with a high software contribution score were consistently ranked at the top (Mscore = 26.41 points, Rankrange = 1-12), while those with a low software contribution score tended to rank at the bottom of the ranking (Mscore = 14.52 points, Rankrange = 16-27). In sharp contrast,

¹Computations have been performed on the supercomputer facilities of the Mésocentre Clermont Auvergne.

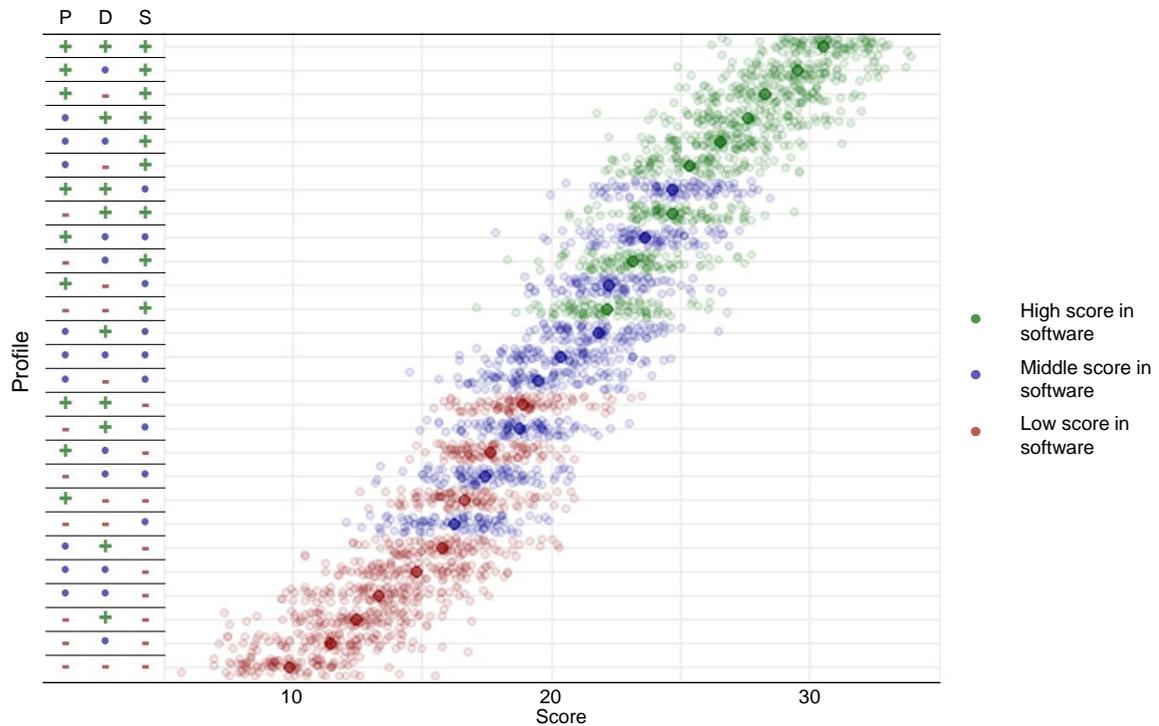


Figure 1

Score obtained by simulated candidates depending on their profiles.

Notes: The x-axis corresponds to the total score and the y-axis corresponds to the 27 simulated profiles, depending on the contributions in papers (P), data (D), and software (S); + = high score, • = middle score, - = low score

candidates with a high contribution score in papers and databases were more evenly ranked across the ranking (Mscore = 23.55 points, Rankrange = 1-20; Mscore = 21.68 points, Rankrange = 1- 25, respectively). Scores and ranks for every type of profile can be found in Figure 1 and Table 1 (the script to simulate candidates are available on OSF: <https://osf.io/y3w4t/>). In addition, we compared the scores of two specific profiles of interest: one with a high score in software but a low score in paper contributions, and one with a high score in paper but a low score in software contributions (both with a mid-score in data). We selected these profiles based on their alignment with the differentiation between a technical contribution to the broader field (e.g., the lavaan package in R: Rosseel, 2012) and an empirical one within a specific domain. On average, the profile with a high score in software obtained 23.15 points, while the profile with high empirical contribution obtained only 17.61 points.

To better assess the impact of each contribution dimension on a candidate's final scores/ranks, we developed a discriminatory index. This index measures the distance in score/rank between the high and low scor-

ers for a given dimension, independent of the other dimensions (i.e., the ability to rank the candidates at the bottom or top of the ladder based on a single dimension). A higher discriminatory index indicates a larger gap in score/rank between the low and high scorers in that specific dimension. This approach allows for a direct comparison of the weights of the three dimensions in the proposed implementation. As expected, given the scales of the dimensions, the discriminatory power of software contributions exceeds that of the other dimensions (see Supplementary Materials).

Taken together, the results of our simulations suggest that software contribution plays a major role in earning points (and ranking) in the current implementation. We acknowledge the importance of this dimension at every step of the psychological inquiry and the scarcity of candidates with such a profile (Yarkoni, 2012). However, we are concerned about the potential unintended consequences of the emphasis on software development, of which we should at least be explicitly aware of. Gärtner et al. (2022) proposed that "any committee may easily adapt these suggestions to the specific necessities in a given hiring promotion process," and we believe that

Simulate Candidates

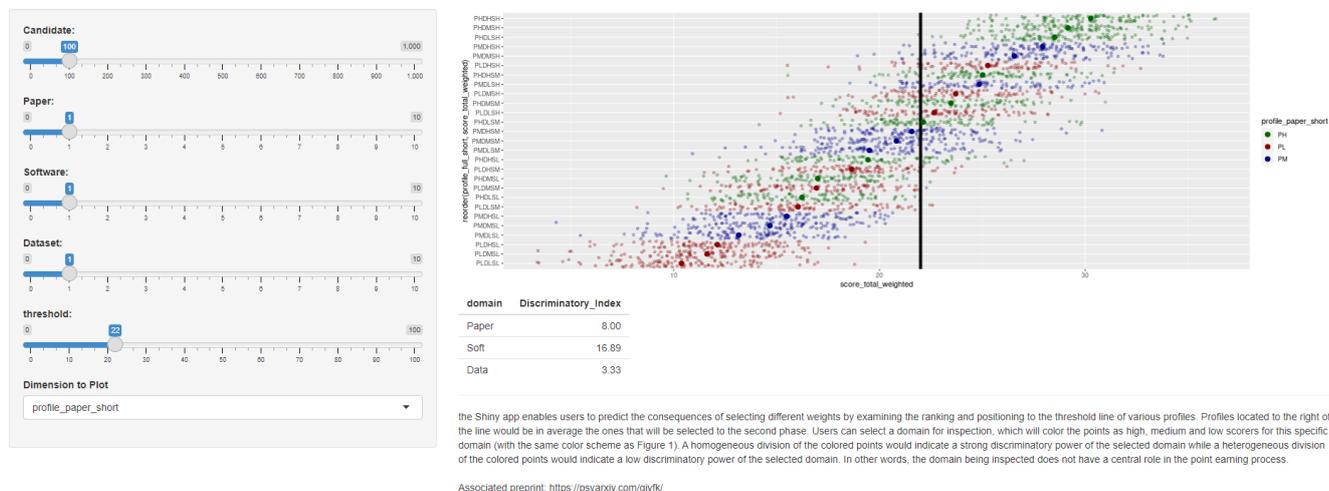


Figure 2

Screenshot of the Shiny app.

Notes: The left-box is the adaptable parameters, in order: number of simulated candidates, the weight of each contribution (paper, software, data), the minimum threshold, and the dimension to visually inspect. The right-top figure corresponds to the score obtained by each candidate grouped by profile. The right bottom table corresponds to the discriminatory index for each contribution.

flexible weights offer a critical solution to address this issue. More broadly, we argue that hiring or promotion processes constitute an opportunity to fill institutions' needs or weaknesses with colleagues possessing specific skills. Thus, recruitment committees could adapt the relative weight of each of the three types of contributions to better match the desired profile. To evaluate the consequences of flexible weights, we developed a Shiny app that simulates plausible candidates and allows for easy comparison between profiles by weighting the three dimensions.

Flexible weights and adaptative hiring process

We have implemented the criteria proposed by Gärtner et al. (2022) in a user-friendly Shiny app (link: <https://tjw41q-victor0auger.shinyapps.io/deploy2/>, see also Figure 2). The app allows users to modify the weights of each contribution and the number of plausible candidates to simulate, using the simulation script (available on OSF) based on predetermined parameters (see Supplementary Materials). The weighted score is computed by multiplying the score on each dimension by the associated weight. Additionally, the user can visually inspect the discriminatory power of the newly weighted dimension by switching between the different dimensions of the profile, as presented in Figure 2.

Furthermore, the Shiny app enables users to predict

the consequences of selecting different weights by examining the ranking and positioning to the threshold line of various profiles. Profiles located to the right of the line would be on average the ones that will be selected for the second phase. Users can select a domain for inspection, which will color the points as high, medium, and low scorers for this specific domain (with the same color scheme as Figure 1). A homogeneous division of the colored points would indicate a strong discriminatory power of the selected domain while a heterogeneous division of the colored points would indicate a low discriminatory power of the selected domain. In other words, the domain being inspected does not have a central role in the point-earning process.

We hope that this user-friendly application will help committees to adopt the approach proposed by Schönbrodt et al. (2022) to improve research quality through hiring and promotion. While the three-dimensional profile approach is a significant contribution to the field, we join Gärtner et al. (2022) in calling for considering more evaluative dimensions on other aspects of scholarly activity (e.g., teaching). In the future, this app could be expanded to allow for the potential weighting of all dimensions of interest.

Discussion

We support the proposition of a more structured, standardized, and transparent hiring process for selecting candidates, as proposed by Schönbrodt et al. (2022), and simulated plausible candidates based on the implementation suggested by Gärtner et al. (2022). Our findings outlined the centrality of the software dimension in candidates' ranking and points earning. While the software dimension is important and represents a significant amount of work, we question whether it should always be the most heavily weighted dimension when hiring researchers in psychology. Our concern is that the current implementation may result in the recruitment of individuals who excel in software development but may not necessarily have the same level of expertise in other important areas, such as producing high-quality research papers or making major contributions to open datasets. These dimensions, which have been shown to significantly contribute to the field and enable theoretical progress, should not be eclipsed by the emphasis on the software dimension (Muthukrishna & Henrich, 2019; Smaldino, 2019).

In addition, we argue that recruitment should aim to fill the institution's needs or weaknesses with new colleagues who possess the desired skills, and that recruitment committees should adapt the relative importance of each contribution based on the searched profile. Following the argument of Schönbrodt et al. (2022) we developed a Shiny App that may assist committees in defining the weights for each contribution. We believe the Shiny App could be a tool to make informed decisions on the relative importance of dimensions of interest. The weights of the different dimensions could also be made public to increase the procedure's transparency and understandability of the committee's decision-making process (Fernandes et al., 2020). Moreover, we encourage the development and operationalization of other scholarly activities (e.g., teaching), to build a more multifaceted and multi-weighted profile. Ultimately, these tools can lead to a more efficient, transparent, and fair hiring process in psychology.

Author Contact

Victor Auger  0000-0002-8895-4591
Nele Claes  0000-0002-1815-0985

Corresponding author: Victor Auger
Address: 17 Rue Paul Collomp, 63000 Clermont-Ferrand, France
Mail: victor.auger@uca.fr

Conflict of Interest and Funding

All authors declare that they have no conflicts of interest. Authors state no funding involved.

Author Contributions

N.C.: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing.
V.A.: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing.

Open Science Practices



This article earned the Open Code badge for making the code openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

- Fernandes, J. D., Sarabipour, S., Smith, C. T., Niemi, N. M., Jadavji, N. M., Kozik, A. J., Holehouse, A. S., Pejaver, V., Symmons, O., Bisson Filho, A. W., & Haage, A. (2020). A survey-based analysis of the academic job market. *eLife*, *9*, e54097. <https://doi.org/10.7554/eLife.54097>
- Gärtner, A., Leising, D., & Schönbrodt, F. (2022). *Responsible research assessment ii: A specific proposal for hiring and promotion in psychology* [PsyArXiv]. <https://doi.org/10.31234/osf.io/5yexm>
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. <https://doi.org/10.1038/s41562-018-0522-1>
- Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of statistical software*, *48*, 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Schönbrodt, F., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., Phan, L. V., Schmitt, M., Scheel, A. M., Schubert, A.-L., Steinberg, U., & Leising, D. (2022). *Responsible research assessment i: Implementing dora for hiring and promotion in psychology* [PsyArXiv]. <https://doi.org/10.31234/osf.io/rgh5b>

- Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, 575(7783), 9–10. <https://doi.org/10.1038/d41586-019-03350-5>
- Yarkoni, T. (2012). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science*, 21(6), 391–397. <https://doi.org/10.1177/0963721412457362>

Appendix

Table 1

Total Score and Rank for each profile

Score	Rank	Profile		
		P	D	S
30.53	1	+	+	+
29.48	2	+	•	+
28.28	3	+	-	+
27.56	4	•	+	+
26.52	5	•	•	+
25.32	6	•	-	+
24.69	7	+	+	•
24.68	8	-	+	+
23.61	9	+	•	•
23.15	10	-	•	+
22.23	11	+	-	•
22.14	12	-	-	+
21.84	13	•	+	•
20.37	14	•	•	•
19.51	15	•	-	•
18.90	16	+	+	-
18.74	17	-	+	•
17.61	18	+	•	-
17.41	19	-	•	•
16.64	20	+	-	-
16.21	21	-	-	•
15.76	22	•	+	-
14.76	23	•	•	-
13.31	24	•	•	-
12.43	25	-	+	-
11.45	26	-	•	-
9.90	27	-	-	-

Note. Data generated based on the simulations of 100 candidates. P = paper contribution, D = data contribution, S = software contribution, + = high score, • = middle score, - = low score