Appendix A

This appendix provides a numerical example concerning what happens if the method assesses a model having missing variables. We created a simplified experiment for the sake of clarity. In the experiment, we assume that thousand real agents take a decision in a deterministic way based on a linear combination of five variables with fixed coefficients. We prepared two very simple models for the validation procedure. In the first, the artificial agents use the same variables and the correct coefficients. In the second, the artificial agents use a truncated number of variables (the first three in the provided file)[1] missing the last two. It uses the right coefficients for the three variables. Then, we obtained the models' outputs. For both outputs, we applied the overall structure of the validation method. We placed real and artificial instances in a latent space[2] and found the optimal number of clusters thanks to the ASW algorithm, which was twelve. Then, we divided the individuals into twelve clusters and obtained the quantity of real and artificial ones as in Table 2.

	Without Missing Variables (Model 1)			With Missing Variables (Model 2)		
Clusters	Artificial Agents	Real Individuals	Score	Artificial Agents	Real Individuals	Score
Cluster 1	109	109	0	103	122	0.08
Cluster 2	84	84	0	100	81	0.10
Cluster 3	78	78	0	83	113	0.10
Cluster 4	105	105	0	99	126	0.12
Cluster 5	68	68	0	80	52	0.21
Cluster 6	92	82	0	86	91	0.02
Cluster 7	75	75	0	94	97	0.01
Cluster 8	92	92	0	57	41	0.16
Cluster 9	85	85	0	83	73	0.06
Cluster 10	53	53	0	70	62	0.06
Cluster 11	74	74	0	65	77	0.08
Cluster 12	85	85	0	80	65	0.10
Overall	**1000**	**1000**	**0**	**1000**	**1000**	**0.09**

Table 2: Quantity of Artificial Agents and Real Instances

As shown in Table 2, the first model produces a perfect balance; the quantity of real and artificial individuals are equal among clusters since the model isolates and considers all required variables. That is the extreme case, but we used it as an example for lucidity. The second model misses two variables, so there are deviations between ex-post behaviors of real and artificial individuals. Therefore, the balance of real and artificial individuals is not as good as the first model. Based on the results in Table 2, the indicator produced two scores for both models. We judged these results thanks to having their statistical distributions and got goodness-of-validation percentages as in Table 3.

Without Missing Variables (Model 1)		With Missing Variables (Model 2)	
Indicator's Score	Goodness-of-validation	Indicator's Score	Goodness-of-validation
0	0.0%	0.09	0.032%

Table 3: The scores and their judgments by statistical distributions

The first model (the extreme case) got zero from the indicator since all the artificial individuals behave similarly to real ones. Thus, it is validated at zero percent since there is no better model for producing a better result. The second model could obtain the score of 0.09961, which is worse than 0.03216% of all

[1] We provide the variables along with the decisions of real and artificial individuals in an excel file on this link: `https://drive.switch.ch/index.php/s/4qiWLhnz8YNaxla`)

[2] We provide a created latent space in an excel file for the sake of clarity: `https://drive.switch.ch/index.php/s/QcE1OMUC2j8ThuL`

possible scores that indicator could produce. The numerical example shows us that missing variables could turn out to have a lower validation score. In other terms, the indicator is sensitive to the misalignment between the model and the empirical world arising from missing variables and produces lower validation for models for which this misalignment is larger. A systematic enquiry into the effects of biases in the coefficients, the number and features of missing variables is, needless to say, beyond the scope of this paper.

Appendix B

In this appendix, we develop a numerical example of how the indicator (see Eq. 4) works and how we can judge the obtained score thanks to the statistical distribution of all possible scores, for the sake of readers. Let's assume that we have 50 artificial and 50 real agents. As explained in Section 2, the number of real and artificial individuals must be equal, which is the initial condition to get a validation score through the method. Let's assume that we have two configurations as examples for the sake of clarity; the optimal number of clusters is obtained as three and five, respectively. For each configuration, we assume three situations; perfect balance, fair balance, weak balance. For each situation, we created the fictitious real-artificial number of individuals among clusters as illustrated in Table 4 to employ the indicator and calculate the score. According to ex-post behavior, which the mDGP generates, one could technically end up with all these situations[3].

Configurations and Situations				
Configurations	Clusters	Perfect Balance	Fair Balance	Weak Balance
3-Clusters	Cluster 1	25 R - 25 A	30 R - 15 A	40 R - 10 A
	Cluster 2	15 R - 15 A	15 R - 25 A	15 R - 25 A
	Cluster 3	10 R - 10 A	5 R - 10 A	5 R - 10 A
	Total	50 R - 50 A	50 R - 50	50 R - 50 A
5-Clusters	Cluster 1	15 R - 15 A	20 R - 12 A	30 R - 5 A
	Cluster 2	10 R - 10 A	8 R - 18 A	5 R - 15 A
	Cluster 3	12 R - 12 A	5 R - 12 A	3 R - 18 A
	Cluster 4	8 R - 8 A	10 R - 3 A	10 R - 2 A
	Cluster 5	5 R - 5 A	7 R - 5 A	2 R - 10 A
	Total	50 R - 50 A	50 R - 50	50 R - 50 A

Table 4: The number of real (R) and artificial (A) individuals in the clusters according to the assumed configurations and situations

Based on the balances among clusters as depicted in Table 4, the indicator in Eq. 4 would create the following scores as illustrated in Table 5 below.

	Situations		
Configuration	Perfect Balance	Fair Balance	Weak Balance
3-Clusters	0	0.3055	0.5888
5-Clusters	0	0.3530	0.6523

Table 5: The scores generated by the indicator

[3]For both configurations, we created state spaces without quantization (i.e., with the quantum size of one). The exhaustive list of state spaces can be downloaded via: `https://drive.switch.ch/index.php/s/demnJYBRJ8xgyje` and `https://drive.switch.ch/index.php/s/uQYgktIiz8XBQ09`

Thanks to the computed all possible scores, we could easily judge the scores in Table 5. We sum up the number of cases that lead to better scores than the specific score and divide it by the total number of cases. So, we could get p-value alike percentages, which are represented in Table 6.

The reasons for computing the statistical distribution of all possible scores to judge a specific score is the following: First, the scores that the indicator could generate are not evenly distributed (see Figure 7). Second, they are sensitive to the number of clusters. Accordingly, the percentage value enables comparability across models, which might have different numbers of clusters or agents. Similarly, it allows to compare how well a model is validated in two time-space contexts (e.g., where the model is better validated in two countries or when the model is better validated across different years). Third, as a percentage and not absolute value is similar to a p-value; thus, the goodness-of-validation can be assessed using very well-established conventional thresholds (e.g., 1% and 5%).

	Situations		
Configuration	Perfect Balance	Fair Balance	Weak Balance
3-Clusters	0	0.0797	0.5163
5-Clusters	0	0.1222	0.6437

Table 6: Percentage values

An example for the interpretation of a particular score and its judgment would be as following: A model, who could achive the score of 0.3055 (fair balance in Table 5) with the given inputs and the found optimal number of clusters (3 clusters for this example), performs worse than 0.0797% of the all possible scores. We interpret the results in this way that this model is validated by the threshold at 0.0797%.

Appendix C

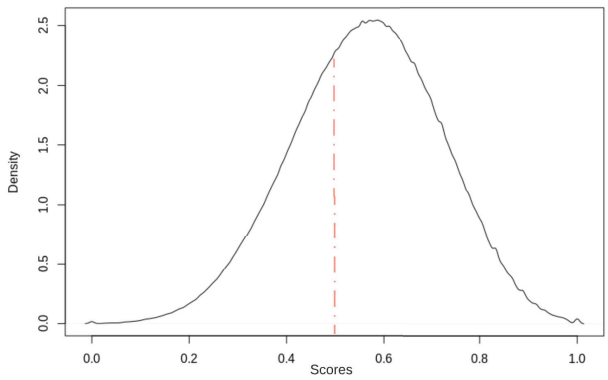

Fig. 8: Density distribution of all possible validation scores with 300 quantum size (instead of 150 in Fig. 7)