**Data Cleaning Process for ICSMP data**

In what follows, we used Yamada et al. (2021) as a blueprint.

We provide all raw datasets that were provided to the ICSMP Consortium by national teams, dominantly in .csv format. These data are openly shared in a folder called Raw data at [https://osf.io/dqmut](https://osf.io/dqmut/) (country data files.zip).

To merge the raw data, we had to introduce minor modifications. We renamed columns to match across data sets, reordered variables alphabetically, and standardized variable labels. Furthermore, all missing values and values denoting the absence of a response were converted to *NA*. Finally, when our statistical software (R & RStudio) produced a warning about the ambiguous format of dates (*start\_date, end\_date, record\_date*), we manually specified the format and transformed variables into dates. The ensuing dataset (ICSMP\_raw\_data.csv) is openly shared in a folder called Raw data merged at https://osf.io/5nqa7/.

At a second stage, we introduced multiple modifications to clean the data for research. Some modifications were introduced to every national data set, while others were introduced only to specific data sets. To each national data set:

* We checked the format of attention check (*attcheck*) and recoded its contents into pass (1) or fail (0).
* Participants’ responses to generosity items (*generosity1 - generosity3*), in case of not summing up to 100 (in data from Croatia, Serbia, South Africa and Latin American countries), were recoded to NA.
* We checked the coding of CRT (*CRT1 - CRT3*) and recoded the responses into intuitive (2), correct (1), and incorrect (0).
* We converted the number of children (*children*) into a variable with a fixed range from zero to ten or more.
* We recoded all dates (*start\_date, end\_date, record\_date*) in the ISO 8601 format (yyyy-mm-dd).
* The age of participants (*age*) who reported being older than 100 was recoded as 100.
* We excluded all participants younger than 18 and duplicated participants. In case multiple participants were recorded with identical inputs within a national data base (except for variables containing dates), only the first input was retained.

Additionally, we compared the distributions of responses per variable across data sets, which pointed out some differences between national data sets. Therefore, we introduced the following changes only to the national data bases where deviations occurred:

* Sex (*sex1*) was recoded in data from Austria, New Zealand to match the numeric inputs in the remaining data sets. In data sets from India1, Canada\_french and Canada\_english, Bulgaria, Germany, and Italy1, responses were provided as text and were recoded into appropriate numerical values to match the coding in other national data sets.
* We checked the patterns of relationships between ladder of socio-economic status (*ladder*) and other relevant variables and recoded it by adding 1 in data from Latvia, Russia and Switzerland. In data from Denmark, Latin American countries, Bangladesh, Romania, India and India2, Canada\_french and Canada\_english, Israel, Bulgaria, Germany, Serbia, Italy and 2, Pakistan, Ireland, China and Ghana, this variable had been inverse-coded and was recoded into 1-11 scale.
* Urbanization profile of participants place of residence (*urban*) was recoded in data from Sweden, Poland, Pakistan, and Ghana to match the numerical values present in remaining national data sets. In data from India1, Italy1, Ireland, and Germany, *urban* was provided as textual variable and was recoded into appropriate numeric values. This variable was missing in the data from Canada\_french and Canada\_english, Bulgaria, where it was coded as NA.
* Data on the number of children (*children*) were not reported in the data set from Ghana, and was coded as NA.
* Duration (*duration*) was not reported in data from Ghana, Russia, Romania, India and India2, South Africa, Bulgaria and Italy2, and was recoded as NA. Duration was computed by comparing start and end time in data sets from South Korea and China and converting the values in seconds.
* Contact variables (*contact1* – *contact5*) were measured on a 0-100 scale in Denmark, Romania, and Romania2, South Africa, and Colombia. The values were divided by 10 to match the original scale. Decimal estimates were found in the data from Taiwan and were rounded to nearest integer. The scale of contact items in Switzerland ranged from 1-11, so 1 was subtracted from each score.
* All five hygiene (*hygiene1 – hygiene5*) and five policy support (*psupport1 – psupport5*) variables were measured on a 0-100 scale in Romania, and Romania2, South Africa, and Colombia. This was also found in data from North Macedonia (item *psupport5*). The values were divided by 10 to match the original scale. Decimal estimates were found in the data from Taiwan (item *hygiene1*) and were rounded to nearest integer. The scale of hygiene and policy support items in Switzerland ranged from 1-11, so 1 was subtracted from each score.
* All the remaining variables in data from Switzerland that were measured on a 1-11 scale were recoded by subtracting 1 from response values to match the range in other data sets.
* In the estimates of risk perception (*riskperc1* and *riskperc2*) in Latin American countries, decimal commas were replaced by decimal points. Values below 1 were found in data from Bulgaria and were multiplied by 100. Additionally, all *riskperc1* values in the data set from Bulgaria that exceeded 100 were recoded as 100. Risk perception in Serbia and Italy2 was measured on a 0-10 scale, thus we multiplied the scores by 10 to match the range present in other national data sets.
* *Know\_tested\_positive*, tested\_positive, *marital1*, and *employ\_status* were provided as text in data sets from Ireland, Italy1, Germany, Bulgaria, Canada\_french, Canada\_english, India1 and were recoded into appropriate numeric values.
* In data from Morocco and Serbia, the values of *know\_tested\_positive* and *tested\_positive* were inverted, while in data from Pakistan responses were textual. In data from Israel, values ranged from 0 to 1. In all cases, initial values were recoded appropriately to match the scale present in the remaining national data sets.
* Responses to moral circle (*moral\_circle*) were provided as a textual variable in data from India1 and was recoded into appropriate numeric values.
* Collective narcissism (*cnarc1 – cnarc3*), conspiracy theories (*ctheory1 – ctheory4*) and optimism (*optim1* and *optim2*) items were measured on a 0-100 scale in data from North Macedonia and were divided by 10 to match their scales in other national data sets.
* Variable *ctheory4* was missing from the Latvian data and its values were coded as NA.
* Happiness (*happy*) and satisfaction with life ladder (*slf\_ladder*) were measured on a 0-100 scale in South Africa and were divided by 10 to match the data from other national data sets.
* Marital and employment status (*marital1* and *employ\_status1*) were recoded in data from Pakistan to match the codes in other national data sets.
* Different CRT (*CRT1* – *CRT3*) items were used in some countries (Latin American countries, Nepal, and Romania). The variables were recoded to match the recoded CRT values in other national data bases.

None of the applied changes interfered in the interpretation of the data, their purpose was to provide the data in a standardized format across national data sets.

The final database with all the included modification was saved as cleaned (ICSMP\_cleaned\_data.csv & ICSMP\_cleaned\_data.sav) at <https://osf.io/ypkrc/> and <https://osf.io/8tyj9/>, respectively.

**Reference**

Yamada, Y., Ćepulić, DB., Coll-Martín, T. *et al.* COVIDiSTRESS Global Survey dataset on psychological and behavioural consequences of the COVID-19 outbreak. *Sci Data* **8,**3 (2021). https://doi.org/10.1038/s41597-020-00784-9