DOCTORAL (PHD) DISSERTATION

Nándor Hajdú

Exploring Contextual Determinants of Human Choices

2024

EÖTVÖS LORÁND UNIVERSITY
FACULTY OF EDUCATION AND PSYCHOLOGY

Nándor Hajdú

Exploring Contextual Determinants of Human Choices

DOI: 10.15476/ELTE.2024.090

Doctoral School of Psychology
Head of Doctoral School: prof. dr. Róbert Urbán

Behavioral Psychology Program
Head of the Program: prof. dr. Anna Veres-Székely

Supervisors: prof. dr. Balázs Zoltán Aczél, dr. Barnabás Imre Szászi

Budapest, 2024

Contents

Acknowledgment

I would like to extend my heartfelt gratitude to Balazs Aczel and Barnabas Szaszi for their mentorship, patience, and invaluable insights. They not only provided academic guidance but also supported my professional and personal growth. I am immensely thankful to my parents, whose faith in me gave me the strength to overcome the challenges that inevitably arise during a doctoral journey. I am grateful for the members of the Metascience Lab and the Behavioral Science Lab, as well as the members of the Affective Psychology Department at ELTE for their support and valuable contributions to my research.

Szeretném kifejezni szívből jövő hálámat Aczél Balázsnak és Szászi Barnabásnak mentorálásukért, türelmükért és felbecsülhetetlen értékű meglátásaikért. Nemcsak tudományos útmutatást nyújtottak, hanem támogatták szakmai és személyes fejlődésemet is. Mérhetetlenül hálás vagyok szüleimnek, akiknek a belém vetett hite adott erőt ahhoz, hogy leküzdjem a doktori út során elkerülhetetlenül felmerülő kihívásokat. Hálás vagyok az ELTE Metatudomány Labor és a Viselkedéstudomány Labor, valamint az Affektív Pszichológia Tanszékének tagjainak a támogatásukért és a kutatásomhoz való értékes hozzájárulásukért.

Introduction

The recognition that many scientific findings may be false has led to reforms in various scientific disciplines, including psychology. Most current solutions that try to reduce the number of false findings aim to improve theory-testing research practices, such as raising awareness of questionable research practices (Wicherts et al., 2016) and promoting preregistration and replication (Nosek et al., 2015). The open science community has been instrumental in advocating for these solutions (Armeni et al., 2021). However, the increasing number of adequately powered, preregistered replication studies has highlighted the reality that many hypotheses that were previously thought to be supported do not survive thorough testing (Scheel et al., 2021). Some researchers suggest a lack of "good theories" as another potential explanation for the replication crisis (Green, 2021). This lack of good theories cannot be resolved through purely confirmatory research, as theory formation necessitates exploratory research. This doctoral dissertation aims to accentuate the importance of exploratory research and show how a specific set of methods, called machine learning methods can be used to create explorative research that can inform theory construction. First, we will describe the empirical research cycle, and how exploratory research fits in it, as a form of inductive research.

The empirical cycle, which outlines the cumulative production of knowledge through scientific research, frames the epistemological function of exploratory research (De Groot & Spiekerman, 2020). This cycle comprises two phases (Wagenmakers et al., 2018): the confirmatory phase and the exploratory phase. In the former, theories give rise to hypotheses that are tested using empirical data. In the latter, observed patterns are consolidated into working theories that can then be tested. This dissertation argues that explorative studies are an important part of research, and that machine learning methods can be used to great effect in order to make explorative research in psychology more informative, and more helpful in theory building. But what is the problem with scientific theory generation in psychology?

Theory generation in psychology

Unlike other scientific fields, such as theoretical physics, where scientists collaborate to generate theories, the field of psychology does not have a comprehensive program for theory construction. Mischel (2008) described this as the *toothbrush problem*:

"Psychologists treat other peoples' theories like toothbrushes — no self-respecting person wants to use anyone else's" (p. 1). In psychology, theories are often the work of (small groups of) individuals. Therefore, there may not be a shortage of theories *per se*, but there is a need for a concerted program of theory development (Borsboom et al., 2021). The lack of this concerted program impedes advancement of the field towards understanding the human psyche in at least three ways (Borsboom et al., 2021). First, because we do not have a firm understanding of how various phenomena relate to one another and whether or not phenomena come from the same underlying principles, it increases the risk of continually reinventing the wheel (Kruglanski, 2001; Vallacher & Nowak, 1997). Second, it is impossible to pinpoint the best interventions for transforming a system in the desired direction without theories that account for causal relationships in a system. For instance, a clearly defined theory of depression would be extremely helpful in creating clinical therapies that are more successful (Borsboom, 2017; Cramer et al., 2016). Third, when developing new studies, we often lack direction without theories. There are advancements in the formalization of theory construction, for instance, the Theory Construction Methodology proposed by Borsboom et al. (2021): they propose that there are at least five, sequential steps of theory construction: (1) identifying relevant phenomena, (2) formulating a prototheory, (3) developing a formal model, (4) checking the adequacy of the formal model, and (5) evaluating the overall worth of the constructed theory. The first two of these steps can be viewed as formalized steps of inductive research.

Inductive research generates working theories from patterns observed in data through qualitative or quantitative exploratory data analysis. Induction infers general principles from specific observations, but the same patterns potentially can be explained by multiple theories. Despite this, induction is a suitable method for generating testable ideas, provided that they are subjected to confirmatory testing. Exploratory research differs from confirmatory research in whether the goal is to test hypotheses or generate them, as well as the flexibility in data analysis; confirmatory research is undermined by researcher degrees of freedom (Wicherts et al., 2016). In ideal circumstances, confirmatory research requires that the hypothesis test be explicitly specified, with only one way to evaluate it (Peikert et al., 2021). Allowing too much flexibility can effectively render a confirmatory study exploratory (Van Lissa, 2022). In contrast, exploratory research entails considerable creativity, with a large number of possibilities for exploring data. A significant portion of the scientific psychological literature is unintentionally exploratory (Van Lissa, 2022),

due to the fact that many psychological theories lack the specificity required to derive testable hypotheses. Confirmatory research, which relies on testing theories through hypotheses, requires strong and detailed theories that are hard to vary: “They explain what they are supposed to explain, they are consistent with other good theories, and they are not easily adaptable to explain anything” (Szollosi & Donkin, 2021, p. 1). In contrast, most psychological theories lack these key characteristics and are therefore unable to generate specific, testable hypotheses. Some theories are so flexible that they can accommodate contradictory evidence without requiring modification, while others are underdetermined to the extent that they cannot be tested at all (Scheel, 2022). As a result, exploratory research has become a necessary tool for generating ideas and working theories from patterns observed in data through qualitative or quantitative exploratory data analysis. Theory formation within psychology presents unique challenges, characterized by fragmentation and ambiguity. The decentralized nature of theory construction inhibits the development of cohesive theoretical frameworks, hindering progress in understanding human behavior. Addressing this challenge requires the introduction of methods that can navigate the complexity of behavioral phenomena and facilitate theory building grounded in empirical evidence. Particularly, this dissertation focuses on the contextual factors of choices in simple choice situations, and how machine learning methods can improve the quality of the explorative studies in this area. However, before we describe how machine learning methods can aid us, we need to discuss the problems in the area of choice architecture and behavioral interventions that, in our opinion, machine learning can help with.

Behavioral interventions and heterogeneity

Theory formation is especially difficult in the case of behavioral intervention research compared to other (bordering) fields of behavioral science or psychology, because many of the detected effects are context-dependent. Some behavior scientists hope to see the coming of a so-called heterogeneity revolution (Bryan et al., 2021), that would bring about a new paradigm of intervention research. In this new paradigm, (1) intervention effects are always thought of as being context dependent; (2) intervention effects that overlook or deemphasize heterogeneity are regarded with skepticism; and (3) there is an understanding that even in cases when there is no type-I error, effect estimates across replications are expected to vary (Bryan et al., 2021). This also means that the criteria we

base our judgment of whether a replication is considered successful or unsuccessful has to be redefined. In their hopes, the implementation and widespread use of these practices would lead to (1) increased attentiveness to the sources of heterogeneity in treatment effects, even in the hypothesis generation phase; (2) emphasis on measurement of research context and characterization of samples, from a heterogeneity viewpoint; (3) statistical methods to detect unhypothesized sources of heterogeneity; and (4) the reduction of the costs of field data collection while keeping the quality of the samples high, with a shared, collaborative infrastructure (Bryan et al., 2021). Our hope is that this dissertation contributes to the first three of these points. The replication crisis has prompted a reevaluation of research practices within behavioral science, revealing limitations in the ability to explain empirical findings with well-founded theories. This gap underscores the complexity inherent in behavioral phenomena and the challenges researchers face in elucidating underlying mechanisms. As such, there is a growing emphasis on methodological approaches that can accommodate this complexity and facilitate a deeper understanding of human behavior.

There exists a common statistical method in order to explain behavior (Agrawal et al., 2020). It starts by the identification of all potential factors that might influence an individual's decisions, and then the construction of a model based on these factors. This involves assessing the statistical significance of each factor, or an overall model measure that balances complexity, like the Akaike Information Criterion (AIC). This helps researchers find a model that strikes the right balance between complexity and accuracy. However, there's a drawback when applying this method to large datasets. In big samples, even smaller effects can achieve statistical significance. As a result, when dealing with sizable datasets, this approach tends to prefer more complex models, even if the improvement in predictive accuracy per data point is minimal, which makes it challenging to extract meaningful insights from the data (Agrawal et al., 2020). Although a point can be made that in reality, there are a vast number of small effects, and a model that encompasses these is the closest to the real phenomenon, making the decision of how to construct these models based purely on statistical significance and information criteria is ill-advised. Instead, examining the combination of the information criteria as well as the prediction accuracy lead to models with more meaningful insights. A more substantial criticism of the previously mentioned approach is that it presupposes prior knowledge of the relevant influencing factors. The question goes beyond simply assessing the

importance of various factors – it involves finding these factors in the first place. Machine learning methods, however, have the potential to enable more rigorous, rule-governed exploration and, by extension, to advance theory formation in psychology.

What is machine learning?

Machine learning is the collective name for data processing processes in which an algorithm performs an optimization process to create a model that can be used to make predictions (Jordan & Mitchell, 2015). These methods hold promise in addressing the complexities of behavioral science. By leveraging machine learning algorithms, researchers can extract patterns from complex datasets and uncover latent structures within behavioral phenomena. This computational approach offers a systematic framework for exploring the multidimensional nature of human behavior. For a more comprehensive introduction to machine learning, consult the work of Hajdu et al. (2023), upon which the sections of this thesis about machine learning are built.The advantages of machine learning over the statistical procedures most commonly used in psychology are that (1) it can test research questions about the accuracy of prediction, (2) it can help provide more accurate and robust estimates (Yarkoni & Westfall, 2017), (3) its results are closer to the needs that lay people and market actors have for science, (4) and it can inform theory building in ways other than hypothesis testing (Hajdu et al., 2023). (1) Machine learning tools can also be applied to address questions regarding the role of each factor within a given theory or model under study. For instance, instead of merely inquiring whether there exists an association between social media use and depression, we can explore which behavioral factors, ranging from nutrition and work stress to leisure activities, are most indicative of someone potentially experiencing depression. Furthermore, we can assess how effectively these factors can be taken into account to estimate the probability of an individual experiencing depression. When we try to estimate the prediction accuracy of our model without splitting our data into training, test, and validation sets, our estimation will be inflated. (2) The latter question could be addressed through traditional statistical methods; however, machine learning offers the advantage of mitigating the risk of overfitting models, thereby enhancing the generalizability of our results. Machine learning models allow for more precise and robust estimations based on new data. Continuing with the previous example, suppose we utilize a machine learning model to analyze new patient data. This approach will yield more

accurate and generalizable estimates compared to not addressing overfitting. These methods, which help prevent overfitting and ensure model robustness, will be discussed in more detail later in this communication. (3) The role of science involves both describing and explaining phenomena, as well as predicting outcomes based on theories. While the latter aspect is less emphasized in psychology, it remains a crucial element in understanding mental phenomena. Those who approach psychology from a non-scientific perspective yet seek to utilize its findings may rightly question why psychological models are challenging to put into practical use. For them, the primary goal is the ability to make accurate predictions about behavior or mental phenomena based on available data—what actions, feelings, or thoughts may be expected when certain characteristics or behaviors are known? Also, while some machine learning models are very complex and hard to interpret for humans, the results of these models are often more easily interpretable than classical statistical models; for example, permutation importance scores, which are changes in the predictive accuracy of the model if a given variable is removed, are easy to interpret. (4) Machine learning models can significantly contribute to theory development by highlighting which factors are predictive of the studied phenomenon and which are not. Factors that prove non-predictive need not be measured or considered in subsequent research, allowing the formulation of hypotheses regarding the predictive elements. Machine learning plays a pivotal role in constructing these models. Moreover, the machine learning framework guarantees the robustness and generalizability of predictions beyond the initial sample, provided that an appropriate sample is used.

Machine learning models can be used to inform theory building (Hajdu et al., 2023). By discerning the non-predictive elements, subsequent research can focus on measuring and considering only the relevant factors, enabling the formulation of hypotheses. Machine learning serves as a valuable tool for constructing these predictive models. Additionally, the machine learning framework ensures the robustness and generalizability of predictions beyond the observed data, provided an appropriate sample is used. In the next section, we describe the three main types of machine learning.

Main types of machine learning

Machine learning encompasses three primary categories: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning. Supervised learning involves

utilizing a labeled dataset to train a model and subsequently evaluating its predictive accuracy on an independent test set. Predictive accuracy, representing the ability of the model to predict properties of new data based on previously observed data, serves as one indicator of model performance. Linear regression, a statistical technique commonly employed in psychology, can also be regarded as a machine learning method since it optimizes a specific parameter, namely the sum of squared errors. The goal is to identify the linear model that minimizes this value among the possible options. In contrast, unsupervised learning does not rely on known correct solutions. Examples of frequently used unsupervised machine learning methods in psychology include principal component analysis and cluster analysis. These techniques facilitate the exploration and identification of underlying patterns and structures within datasets. Reinforcement learning, the third machine learning method, operates by reinforcing desired behaviors and penalizing undesired ones. In this approach, a learning agent is capable of taking actions and learning from feedback received based on the outcomes of those actions. In the following sections, we make a case of why machine learning methods, particularly supervised learning methods, are useful for exploratory analysis.

Typical supervised learning workflow

The typical workflow of defining supervised machine learning models is as follows: first, we select a suitable learning algorithm for the task at hand. If the algorithm requires hyperparameters for fine-tuning, we provide them accordingly. Subsequently, the algorithm undergoes an iterative process using the training data, culminating in the acquisition of optimal model parameters. The resulting model is then evaluated, and if necessary, the previous steps are repeated with adjusted hyperparameters to obtain the most optimal model. Finally, we assess the performance of this model using new data and examine the predictions it generates. In order to make predictions regarding the output variable, it is necessary to construct a mathematical model that represents the phenomenon being studied. In the field of machine learning, the learning algorithm iteratively determines the optimal parameters of the model. For instance, in linear regression, the Ordinary Least Squares algorithm identifies the intercept and slope of the optimal regression line, which describes how the input variables predict the output variable's value. However, all models are simplifications and therefore imperfect

representations of reality. The imperfections, or errors, in the model can be divided into two components.

Bias and variance

The first component is the bias, which arises from the limitations of the algorithm used for modeling, constraining the range of possible solutions and affecting their accuracy. For example, a linear model can only inadequately represent nonlinear relationships or fail to capture them entirely. In such cases, the model attempts to depict the relationships between variables in an overly simplistic manner, resulting in an inability to accurately describe the true nature of the relationships. This phenomenon is known as underfitting. On the other hand, the problem with variance error lies in the model's excessive sensitivity to the variability observed in the sample. This error occurs when the model incorporates noise present in the sample, such as sampling error, leading to excessive complexity. Consequently, the model may explain a high proportion of variance within the sample, but its predictions may not generalize well outside the sample. This situation is referred to as overfitting.

When specifying our workflow, we have to be aware of data leakage, a phenomenon marked by the improper use of external information during model creation, leading to artificially inflated performance metrics. This risk arises when the same dataset is employed for both model training and performance assessment or when unavailable features are incorporated into the model. For example, predicting participants' recall performance on a specific test should not involve data from a subsequent observation of recall performance on another test, as it could compromise prediction accuracy. Models afflicted by data leakage tend to excel on training data but perform poorly on previously unseen data. There are several causes and forms of data leakage. (1) Leakage from the future occurs when the model is trained on data that includes information from the future that would not be available in a real-world scenario. For example, including the target variable from a future time period in your training data can lead to unrealistic performance. (2) Data preprocessing, such as feature scaling or imputation, can introduce leakage if it's done without proper consideration of the data separation between training and testing sets. For instance, scaling or imputing data based on the entire dataset, including the testing set, can introduce information from the testing set into the training process. Using global statistics, such as the mean or standard deviation of a feature across

the entire dataset, can introduce information from the testing set into the training process. (3) If data is not properly split into training and testing sets, meaning that some data points are in both the training and test sets, it can lead to data leakage. Leakage occurs when information from the testing data is inadvertently used in the training process.

When constructing models, it is crucial to strike a balance between bias and variance. One must be mindful of the limitations inherent in the chosen learning algorithm, as they may predispose the creation of a suboptimal model characterized by high bias or high variance. While addressing bias often requires selecting a different learning algorithm, there are several techniques available for detecting and mitigating overfitting, which are typically integral parts of the machine learning workflow. While striking a balance between bias and variance is a large benefit of machine learning methods, there are other benefits, as well. One of these is that various variable selection methods can be applied in order to find the variables that lead to the most accurate predictions.

Variable selection

In some cases, it is advisable to omit certain variables to yield more concise and interpretable models. This is especially important when the number of variables would make the model very difficult to interpret. Unfortunately, the pursuit of simplicity can result in models that fail to grasp real causal relationships. This pursuit led to deleterious practices in variable selection, such as stepwise regression analysis, gaining prevalence in social sciences (Gigerenzer, 2004; Thiese et al., 2015). Stepwise regression methods have been found to result in too high estimated R^2 values, too small standard errors, too low p-values and potential collinearity problems (Harrell et al., 2001). Based on Monte Carlo simulations, stepwise regression might omit variables that have a real causal relationship with the dependent variable, while other, not as important variables are shown to have a significant effect (Smith, 2018). In machine learning, where the number of independent variables can be very high, it is also very important to find an effective and appropriate method to choose the variables we need and ignore those we do not. The area and practice of choosing which variables or so called features to include is called feature selection. Feature selection serves to find the most important and informative features that are capable of encapsulating the inherent data patterns. Within the machine

learning paradigm, feature selection plays a pivotal role, contributing to the management of model complexity. Feature selection aims to curtail data dimensionality by removing redundant or irrelevant attributes, thereby potentially enhancing model performance and interpretability.

Broadly, these selection techniques are categorized into three groups: (1) filter methods, (2) feature subset selection (wrapper) methods, and (3) embedded methods (Jovic et al., 2015). Filter methods leverage statistical feature properties like correlation or shared information content, operating independently of the model. Features surpassing a predefined threshold score are kept, while those falling below are discarded. This subset can subsequently be used as an input to the chosen algorithm.

Filter methods offer computational efficiency, a virtue distinguishing them from other approaches and rendering them adaptable to high-dimensional datasets. However, they are model-agnostic, a feature that can limit their efficacy. In contrast, feature subset selection methods harness a model to gauge prediction performance across various feature sets, opting for the one yielding superior outcomes. This approach considers feature interactions and their interplay with the model, potentially resulting in heightened accuracy. Yet, these methods tend to be computationally intensive, and the selected feature sets may not be universally applicable across different algorithms. For instance, features chosen optimally for a linear regression model via feature subset selection may yield subpar results in a random forest model. This disparity, though not eliminated in filtering methods, is less pronounced.

Embedded methods seamlessly integrate feature selection into the model's algorithm. During training, the model adapts internal parameters to assign optimal weights to each feature, optimizing accuracy. Consequently, embedded methods consolidate feature selection and model construction into a unified step (Guyon & Elisseeff, 2003). Notably, these methods encompass feature selection via regularization, as observed in LASSO models previously mentioned.

Features exclusive to the training phase that are not available during testing, should be excluded from the model. After the variable selection process, we have the data we need, but how can we tell which variable has the greatest effect on the prediction accuracy?

Variable importance

After we know which variables to include in our model, the next step is to see which of the predictors have an effect. Machine learning allows for comprehensive inclusion of relevant predictors, which is important for theory formation (Van Lissa, 2022). To compare the importance of predictors within the same model, we need to determine which predictors are most strongly associated with the outcome. This is not possible if each study examines only one piece of the puzzle. Machine learning offers a more holistic approach by identifying the most strongly associated predictors with the outcome of interest and guiding exploration by existing theories or including factors not yet represented in theory. This variable importance is how we translate the knowledge we get from our models to theories. How we do this is by checking whether there is a congruence with theoretical assumptions about important predictors. Some theorized predictors may be less important, and some undertheorized predictors might be more important than previously thought. This gives additional information to revisit and revise theories and accelerate theory (re)construction.

Dissertation goals

In our current day and age, when computational power is cheap and easy to come by, our opportunities to define and test complex models of theories is greater than ever. We do not have to necessarily oversimplify theories for them to be testable. But, even with the valuable qualities of the machine learning toolkit that allow us to overcome overfitting and data leakage, the potentially highly automated process of finding the most relevant predictors is still only a part of the puzzle of theory construction. As researchers, we still have a role in theory construction, because we still have to consolidate the insights we get from exploratory analyses into theories. If we make the exploratory analysis more streamlined, controlled, and rules-governed, while keeping its exploratory nature intact, we arrive at a set of results that stand on a robust basis. As we navigate the complexities of behavioral science, there is a growing recognition of the importance of methodological innovation in advancing our understanding of human behavior. By embracing the challenges posed by complexity and harnessing the power of machine learning, researchers have the opportunity to chart a new course for behavioral science—one

characterized by rigor, transparency, and a deeper appreciation for the intricacies of human behavior.

Our research detailed in this dissertation is about using mostly supervised, but in some cases, also unsupervised machine learning methods in exploratory analyses. Here we present three research papers that use these methods to great effect, with gradually less information about the choice situation in each research project. We present the efficacy of the machine learning approach in three different levels of data availability (Table 1).

Title	Topic	Goal	Result	Published in
Extending the choice architecture toolbox: The Choice Context Exploration	Predictors of choosing stairs over elevators when going upstairs	To find the potential influencing factors of stairs-elevator choice in a setting in a specific environment, where most of the possible predictors could be measured	The selected influencing factors predicted choice with >90% accuracy	Sage Open
Contextual factors predicting compliance behavior during the COVID-19 pandemic: A machine learning analysis on survey data from 16 countries	Predictors of compliance with COVID-19 regulations	To find the potential influencing factors of compliance in a measurement configuration where the environment (the participant's home) could not be measured fully	The selected influencing factors predicted choice with 62 - 87% accuracy	PLoS ONE

A machine learning analysis of the relationship of demographics and social gathering attendance from 41 countries during pandemic	Demographic predictors of social gathering attendance	To find the potential influencing demographic factors of social gathering attendance, with no other contextual data available	The selected influencing factors predicted choice with 52 - 84% accuracy	Scientific Reports
Applying behavioral interventions in a new context	The role of heterogeneity in behavioral intervention planning	To aid experts in intervention planning raising awareness about heterogeneity	-	Behavioral Science in the Wild

Table 1. Articles and/or book chapters included in the dissertation with their respective main topic, goals, and results.

In the first research paper, we use machine learning tools to explore the reasons why people choose stairs over elevators when going up in a building, or vice versa. We looked for the potential explanatory variables that lead to the most accurate predictions of what people would do. In this situation, the decision was relatively easy to track, and every major theorized potentially influencing factor could be, and was accounted for. In our second research paper presented in this dissertation, we were looking for contextual factors predicting compliance behavior during the COVID-19 pandemic. In this case, the definition and the context of choice is more fuzzy than in the previous article, and fewer predictors were available. In our third paper, we were interested in the prediction of social gathering attendance during the COVID-19 pandemic, based on demographic variables. In this case, we had the least amount of information that could be used for prediction. Finally, after the three research papers, we present a book chapter that summarizes our thoughts on choice context exploration and gives insights on how to plan better interventions, accounting for contextual heterogeneity.

References

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, *117*(16), 8825–8835.

Armeni, K., Brinkman, L., Carlsson, R., Eerland, A., Fijten, R., Fondberg, R., Heininga, V. E., Heunis, S., Koh, W. Q., Masselink, M., Moran, N., Baoill, A. Ó., Sarafoglou, A., Schettino, A., Schwamm, H., Sjoerds, Z., Teperek, M., van den Akker, O. R., van't Veer, A., & Zurita-Milla, R. (2021). Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, *48*(5), 605–611. https://doi.org/10.1093/scipol/scab039

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13.

Borsboom, D., & Markus, K. A. (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement*, *50*(1), 110–114. https://doi.org/10.1111/jedm.12006

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766.

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Cramer, A. O., Van Borkulo, C. D., Giltay, E. J., Van Der Maas, H. L., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic

system. *PloS One*, *11*(12), e0167490.

De Groot, A. D., & Spiekerman, J. A. (2020). *Methodology: Foundations of inference and research in the behavioral sciences* (Vol. 6). Walter de Gruyter GmbH & Co KG.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Green, C. D. (2021). Perhaps Psychology's Replication Crisis is a Theoretical Crisis that is Only Masquerading as a Statistical One. *International Review of Theoretical Psychologies*, *1*(2), Article 2. https://doi.org/10.7146/irtp.v1i2.127764

Guyon, I., & Elisseeff, A. (n.d.). *An Introduction to Variable and Feature Selection*.

Hajdu, N., Szaszi, B., Aczel, B., & Nagy, T. (2023, September 18). Using supervised machine learning methods in psychological research. https://doi.org/10.31234/osf.io/tjkug

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis* (Vol. 608). New York: Springer.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Kruglanski, A. W. (2001). That" vision thing": The state of theory in social and personality psychology at the edge of the new millennium. *Journal of*

Personality and Social Psychology, *80*(6), 871.

Mischel, W. (2008). The Toothbrush Problem. *APS Observer*, *21*. https://www.psychologicalscience.org/observer/the-toothbrush-problem

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Peikert, A., Van Lissa, C. J., & Brandmaier, A. M. (2021). Reproducible Research in R: A Tutorial on How to Do the Same Thing More Than Once. *Psych*, *3*(4), 836–867. https://doi.org/10.3390/psych3040053

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), e2295.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 251524592110074. https://doi.org/10.1177/25152459211007467

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, *5*(1), 1–12.

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, *16*(4), 717–724.

Thiese, M. S., Arnold, Z. C., & Walker, S. D. (2015). The misuse and abuse of statistics in biomedical research. *Biochemia Medica*, 5–11. https://doi.org/10.11613/BM.2015.001

Vallacher, R. R., & Nowak, A. (1997). The emergence of dynamical social psychology.

Psychological Inquiry, *8*(2), 73–99.

Van Lissa, C. J. (2022). Developmental data science: How machine learning can advance theory formation in Developmental Psychology. *Infant and Child Development*, e2370.

Wagenmakers, E.-J., Dutilh, G., & Sarafoglou, A. (2018). The Creativity-Verification Cycle in Psychological Science: New Methods to Combat Old Idols. *Perspectives on Psychological Science*, *13*(4), 418–427. https://doi.org/10.1177/1745691618771357

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01832

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Chapter I.

Hajdu, N., Szaszi, B., Aczel, B. (2023). Extending the choice architecture toolbox: The Choice Context Exploration. *Sage Open*

Nandor Hajdu[ab], Barnabas Szaszi[b], Balazs Aczel[b]

[a]*Doctoral School of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary*

[b]*Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary*

Abstract

The importance of context in behavioral interventions is undeniable, yet few intervention studies begin with a systematic investigation of the contextual factors that influence the behavior in question. This is largely due to the lack of a reliable method for doing so. In recognition of this gap in the field, we have developed a procedure called the Choice Context Exploration that uses machine learning tools to examine the contextual factors that influence a targeted behavior. We demonstrate the steps of Choice Context Exploration using the example of the behavioral choice between using stairs or an elevator. Potential contextual factors were identified by laypeople and experts, and two surveys were created to measure both the behavior and choice, as well as the beliefs of participants. We estimated the effect of contextual factors on participants' behavior and were able to identify the most influential ones in relation to the studied choice. We achieved an accurate prediction of whether participants would choose the stairs or the elevator based on contextual information in 91.43% of cases on previously unseen data. We also found that participants had different beliefs about what influenced their choice in this situation and that they could be divided into different groups based on these beliefs. Our results suggest that the Choice Context Exploration is a useful procedure for collecting and assessing contextual factors in a given choice setting, which can aid in the planning of behavioral interventions by significantly reducing the number of potential interventions that are likely to be effective.

Keywords: Choice Context Exploration, choice architecture; nudge; behavioral interventions; generalizability

Introduction

Consider a scenario in which you want to encourage people to engage in a healthy behavior, such as using the stairs instead of the elevator. The literature offers many examples of different nudge techniques that have been used in various settings with varying degrees of success (Duflo et al., 2011; Silva & John, 2017). How do you decide which nudge technique to use? We argue that researchers cannot make an informed decision about their intervention technique until they have explored the influential contextual factors of the studied choice situations, particularly in behavioral change interventions where the method does not limit choices. This paper presents procedural steps for detecting contextual influences to aid in choice architecture interventions.

Over the past decade, nudge interventions have gained popularity as a method for changing behavior on a large scale. Nudge, as outlined in Thaler and Sunstein's book (2008), is the influencing of behavior by altering choice architecture using relatively inexpensive and non-intrusive methods that take advantage of general cognitive processes and biases. While nudges can be successful in many cases (John et al., 2013), they can also be ineffective (Silva & John, 2017) or have only temporary effects (Brandon et al., 2017).

One potential reason for this inconsistency in results may be that, in the prevalent culture of behavioral intervention research, nudge researchers aim to find all-encompassing effect sizes and do not consider the potential heterogeneity across various contexts (Tipton et al., 2019). As a result, the reasons for why, when, and to what extent interventions work or don't work are often unclear. It has been suggested that instead of trial-and-error assessment of ad hoc interventions in a given context, researchers should focus on advancing theory and exploring moderators (Szaszi et al., 2018). While there are models of behavior that can be used to design field experiments and make the interpretation of results easier and more convincing, they are often lacking in the design of interventions.

However, various toolkits are employed to design behavioral interventions that enable the customization of the intervention to suit specific contexts. One of the most widely used toolkits for designing behavioral interventions is the Behavior Change Wheel (Michie, van Stralen & West, 2011). This framework divides the problem of intervention

planning into three layers. The first layer focuses on identifying the sources of behavior that might be targeted using the COM-B model, which stands for Capability, Opportunity, Motivation, and Behavior. The second layer includes intervention functions to choose from, including Education, Persuasion, Incentivization, Coercion, Training, Enablement, Modeling, Environmental Restructuring, and Restrictions. The third layer contains policy categories that can be used to deliver the intervention. In their book, Atkins, West and Michie (2014) provide a more detailed eight-step guide for using the Behavior Change Wheel.

A commonly used theoretical framework in behavioral intervention design is the Model-based/Model-free framework (Daw et al., 2011; Marteau et al., 2020). This framework treats behavior as a bidirectional interaction between an agent and its environment, where the agent receives information about the environment and acts in a way that maximizes reward. The model includes the agent, a set of possible actions, and a set of action-dependent outcomes. These outcomes are probabilistically associated with rewards and a possible change in the environmental state. The agent may update their behavior in a model-free way, meaning they are influenced by the short-term rewards resulting from their actions, or in a model-based way, where the agent chooses actions with lower short-term rewards but higher long-term rewards. In order to generate useful models of behavior that can be used to explain intervention effects, this framework also requires a thorough understanding of the environment, in addition to the actor and potential rewards. These frameworks give useful insights and methods of collecting contextual information, while also offering substantial flexibility in analyses. This flexibility, however, comes with the price of the possibility of not choosing the most adequate analysis method for the task. We propose that an exhaustive and reproducible exploration of relevant contextual information can be achieved by the use of a combination of machine learning methods. The present paper aims to provide a step-by-step guide for collecting the potential contextual influences of a given environment while showcasing the use of machine learning methods for context exploration.

The focus is on context

The elements of context that constitute an environment can be categorized in various ways for operationalization purposes. Finding an adequate working definition of what context constitutes is challenging, because most studies provide a narrow

conceptualization of context, and simply list contextual determinants of a construct (Nilsen & Bernhardsson, 2019; Rogers, De Brún & McAuliffe, 2020). As the result of their systematic review, Rogers and colleagues defined context "as a multi-dimensional construct encompassing micro, meso and macro level determinants that are pre-existing, dynamic and emergent throughout the implementation process. These factors are inextricably intertwined, incorporating multi-level concepts such as culture, leadership and the availability of resources" (Rogers et al., 2020, pp. 18). Our definition of context is based on the criterion of choosing the source of information within the environment as a prescription. We define context as the physical environment, such as surrounding objects or creatures, the intrapersonal circumstances, such as mental states, and sociocultural environment, such as customs and norms, present at the time of the choice that may affect decisions. We argue that while knowing the context is a necessary, but not sufficient, prerequisite for intervention choice, exploration of potential influencing factors is needed while there is no intervention in effect in order to later model their interactions with the effects of interventions. This way, it is possible to advance the general understanding of how interventions work. Sufficient knowledge of context is crucial. In this sense, exploring contextual information is analogous to taking a patient's medical history in therapy settings; while the therapeutic methods to be used cannot be decided based solely on the medical history, it is still a vital part of the process as it helps in determining further directions.

The role of context in interventions is rarely completely ignored, but its investigation is often not systematic (Szaszi et al., 2018). There are two main sources of information indicating which potential contextual information to consider when planning an intervention: either based on some insight, without empirical data, or on published results. Insights-based contextual information may not necessarily come from the researchers' own experience and opinion; it can also come from the opinions of other experts, collected either through interviews or casual conversations about the topic. Valuable information can also be gathered by convening a demographically diverse focus group and discussing the issue at length (Puchta & Potter, 2004). However, the insights approach has its limitations: its strength depends on the validity and breadth of these insights, which are rarely known or assessed. Empirical contextual information may be obtained from relevant literature: specific interventions, reviews, and meta-analyses. The main limitation of these sources is that the generalizability of the findings is limited to the

context of the original studies, and there are only a few studies exploring the question of generalizability (Szaszi et al., 2018). Furthermore, we cannot possibly know the contextual influences without previously exploring them in depth.

Existing frameworks

There are numerous frameworks and guidelines that can be useful when planning behavioral change interventions. For example, MINDSPACE (Dolan et al., 2010) provides a more in-depth aid, describing nine robust influences on human behavior: Messenger, Incentives, Norms, Defaults, Salience, Priming, Affect, Commitments and Ego, most of which can be context-related. The guide by Ly, Mažar, Zhao and Soman (2013) emphasizes the importance of contextual exploration and provides a useful set of questions about the properties of the decision, such as incentives and cost, sources of information for the individual making the decision, features of the individual's mindset, and various environmental factors that can help in planning interventions. However, no instructions are provided on how to answer these questions. EAST (Algate et al., n.d.) is a purposely simple framework to follow, aimed at policy-makers rather than researchers. It argues that planning and executing a behavioral intervention can be more successful if several attributes of the intervention are set before planning smaller details: it should be Easy, Attractive, Social, and Timely. Another framework, the BASIC approach developed by the iNudgeyou team (Schmidt et al., 2016), offers a guideline for planning interventions with an emphasis on applicability. The first step is Behavioral exploration, which involves collecting data through observations of the target population. While observation is a valuable source of information, it does not necessarily deepen the understanding of the reasons and motivations behind the behavior. Although these frameworks can certainly help in designing nudge interventions, and some of them emphasize the importance of contextual information, they do not provide a standardized method for investigating and measuring these factors. The exact methods for prior assessment are left for the reader to devise. We argue that the lack of thorough and comprehensive exploration of moderators and potential contextual influences is one of the main obstacles in developing extensive theoretical frameworks for large-scale behavioral interventions (Szaszi et al., 2018).

Choice Context Exploration

We developed the Choice Context Exploration, a procedure that aims to help the researcher explore the influential contextual factors in a given choice situation. Our main motivation behind creating the Choice Context Exploration is to define a set of steps that result in accurate predictions of people's choices. The procedure consists of four steps.

Step 1: Collecting Potential Influencing Factors

The goal here is to gather information from diverse sources, including the relevant professional literature, the target population, and experts about what attributes of the context (i.e., factors) might influence the given choice. These factors can refer to the physical attributes of the environment, the nonphysical factors such as social, cultural, or psychological attributes of the target population, as well as the timing of the choice. This step can be achieved through asking experts and laypeople (e.g., through questionnaires, interviews, or focus group discussions) about the potential influencing factors of the behavior in question. This exploration can bring details to the surface that are not available from the literature. In order to create a final list of potential factors, the collection needs to be curated by merging all elements that refer to the same attributes of the context.

Step 2: Quantifying the Influence of Factors

The aim of Step 2 is (a) to measure each contextual factor that potentially influences the choice in question, along with a measurement of the choices themselves, as well as (b) to understand the extent to which each contextual factor contributes to the observed behavior. The data collected can be used to estimate the strength and direction of the relationship between the observed behavior and the gathered contextual factors. When planning interventions, these estimates can be used to predict changes in behavior when these contextual factors change. Collecting behavioral data in natural settings is valuable because people's beliefs about and observations of choice behavior may not fully overlap. By dividing the collected data into training and test sets, it is possible to test the predictive properties of our contextual factors on data not used for model specification.

Step 3: Assessing Beliefs about the Influence of Factors

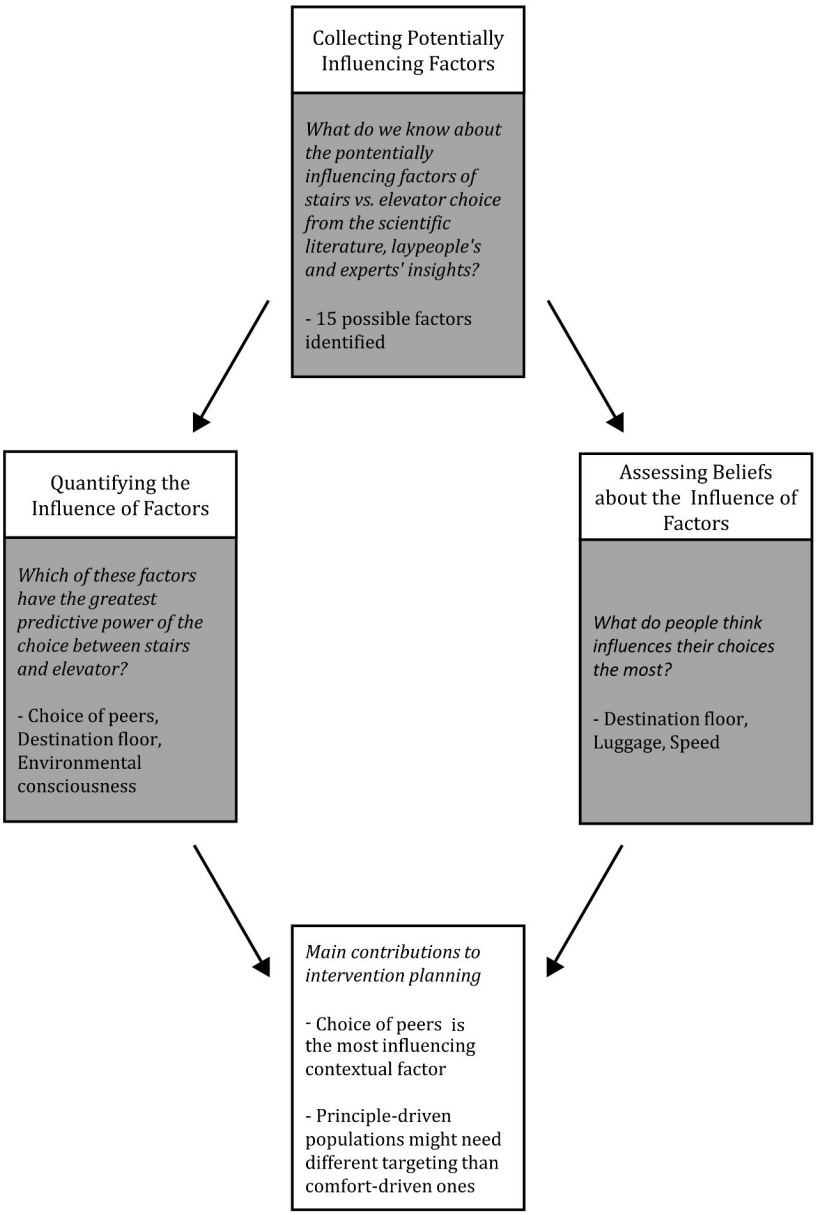
The aim is to measure (a) the extent to which people believe that the factors collected in Step 1 influence their behavior in a choice context and also (b) to explore whether people can be grouped based on their beliefs about what influences their elevator/stair choices, and, if possible, the relative sizes of these groups. First, a survey that allows for quantitative measurements must be created in order to assess the beliefs of individuals. Then, it must be determined whether meaningful clusters of the sample can be formulated based on beliefs of the contextual influences. These clusters provide information about which beliefs occur together, as well as the relative ratios of people with the same thinking about the situation. This can be helpful when planning interventions, as using cues that the most people are sensitive to may potentially have the largest overall impact.

Step 4: Comparative analysis

For those who wish to find out whether people's choice-related beliefs and their corresponding behavior are aligned, we recommend comparing the results of Steps 2 and 3. This analysis can indicate how aware people are of the causes of their choice in a given context, or whether they hold false beliefs about their behavior. It can provide insight into how much we can rely on people's insights in understanding the choice context. It may not be possible to compare models defining the relationships between the contextual factors and behavior with the models describing beliefs about these contextual factors directly. However, the relative order of effect sizes can be contrasted.

By following this procedure, relevant contextual factors can be identified and their effects on the target behavior measured.

Figure 1. Steps and results of the Choice Context Exploration in the case of stairs and elevator use.



Choice Context Exploration in Practice

In this study, we explored the use of Choice Context Exploration in the context of individuals making a decision between using the stairs or elevator. This topic is of interest

because there may be multiple factors that influence people's default choice. Previous research on this topic has produced mixed results, possibly due to the variability of unexplored factors among studies (Bellicha et al., 2015; Jennings et al., 2017). To identify and assess the contextual factors that influence the use of stairs and elevators, we conducted a study with a sample of university students. Using the Choice Context Exploration method, we followed the following steps: (1) we surveyed experts and laypeople to collect potential factors influencing the decision between these two means of movement, (2) we investigated how well the factors identified in step 1 explain individuals' behavior when choosing between the stairs and elevator, (3) we studied participants' beliefs about what influences their choices in this situation, and (4) we compared the results of steps 2 and 3 to determine the degree to which participants' beliefs correspond to their behavior.

Step 1 - Collecting Potential Influencing Factors

In step 1, our goal was to gather a list of potential factors that might influence whether people use the stairs or the elevator. To achieve this, we surveyed a sample of university students and asked experts to provide open-ended responses about the potential factors. The research plan was approved by the local institutional ethical review board.

Method

We randomly selected 500 individuals from the subject pool at our local university in Hungary, which consisted of students who had signed up for a course where they could participate in various studies in exchange for course credits. We were able to successfully recruit 392 of these students, who were eligible if they were at least 18 years old and received course credits as compensation. We asked these participants to list the contextual factors that they believed influenced their own and others' decisions between stairs and elevators.

Secondly, we identified experts by compiling a list of those who had published at least one peer-reviewed research article on the topic of stair usage interventions in the past decade. We asked these experts to list the potential factors that might influence the choice between stairs and elevators. Out of the 47 experts we contacted, seven responded.

We then processed each of the collected responses by one member of our research team, registering new categories for each type of influencing factor that was mentioned. If a newly processed response did not fit into any of the existing categories, a new category was created. Finally, we also reviewed relevant professional literature for additional contextual influencing factors: we searched for papers about interventions that targeted staircase and elevator use. (For the demographics of the respondents and the wording of the surveys, see the Supplementary Materials).

Results

As a result, 16 potential influencing factors were identified: *Appeal of stairs/elevator, Comfort/Laziness, Destination Floor, Elevator availability, Environmental consciousness, Fear of confined spaces and/or technical problems, Fatigue, Importance of Health/Sports, Luggage, Number of people in the elevator, Peer behavior, Physical limitations, Speed, Speed of elevator, Stairs/elevator physical availability, Temperature*. *Appeal* is a factor only suggested by experts and aggregates physical aspects of a staircase that make approaching it a better experience. An example of the mention of *Appeal* would be "the design of the stair should be inviting, open, bright, and ventilated"; another example is "the physical appearance and condition of the stairs (often not well lit, maybe smelly)". *Table 1* shows the percentage of experts and laypeople mentioning each category.

Table 1. Percentages of factor occurrences in free-form text answers of experts and laypeople.

Factor	Mentioned by % of experts (N = 7)	Mentioned by % of laypeople (N = 392)
Appeal of stairs/elevator	71.43	0
Comfort/Laziness	0	78.83
Destination Floor	57.14	68.11

Elevator availability	28.57	20.66
Environmental consciousness	0.00	3.83
Fatigue	14.29	40.05
Fear of confined spaces and/or technical problems	0	33.67
Importance of Health/Sports	14.29	76.28
Luggage/Carrying additional items	14.29	33.93
Number of people in the elevator	0	32.91
Peer behavior	42.86	12.76
Physical limitations	42.86	40.05
Speed	28.57	80.61
Speed of elevator	14.29	0
Stairs/elevator physical availability	85.71	2.81
Temperature	14.29	0

Step 2 - Quantifying the Influence of Factors

In step 2, we collected behavioral and contextual data to assess the extent to which the contextual factors identified in step 1 influence the choice between stairs and elevators. The methods and analysis procedure for Step 2 and Step 3 were pre-registered at https://osf.io/bp265 . Deviations from the original pre-registered procedure are described in the supplementary materials section.

We collected data for step 2 from the same participant pool as for step 3, and participants were randomly assigned to either complete step 2 followed by step 3, or step 3 followed

by step 2. By randomizing the order in which participants completed the surveys, we controlled for the potential influence of realized beliefs on behavior.

Method

Participants were 523 (346 female, 177 male) Hungarian university students (M_{age} = 21.87 years, *SD* = 3.26) who were over 18 years old, recruited via email advertisement and received course credits as compensation. The sample size for this study was chosen based on availability, and the research plan was approved by the local institutional ethical review board. Participants accessed our online questionnaire through a link provided in an email. The questionnaire first asked them to indicate whether they had visited any of the university buildings the previous day, or if they had not visited higher floors that day. They were also asked to report which building they had last been in and whether they had chosen the elevator or stairs to go upstairs.

Next, we asked participants to indicate the parameters of the contextual factors at the time of the choice. For example, for the *Luggage* factor, they were asked to rate on an 11-point Likert-type scale (ranging from "extremely disagree" to "extremely agree") how much they agreed that they were carrying a heavy luggage at the time of the choice. One exception was the *Peers* factor, which was an item with three options, where participants indicated whether they had no peers with them at the time of their choice, or whether their peers opted to use the elevator or stairs.

We only included factors identified in step 1 in the questionnaire, where variability in participants' choices was expected. Therefore, the *Physical limitations* and *Claustrophobia and technical problems* factors were excluded, as participants indicating such physical or mental conditions would have been prevented from making a choice between using the stairs or the elevator. The items of the questionnaire are provided in the supplementary materials.

Participants were asked to indicate any physical or mental conditions that would prevent them from using the stairs or the elevator (e.g., injury, claustrophobia). These factors, *Physical limitations* and *Claustrophobia and technical problems*, were accounted for in the questionnaire. However, we chose not to include data from participants who reported having their options limited by physical or psychological conditions, as there would have been very low variance in their answers.

Participants received the same questionnaire on 10 consecutive weekdays. They were told that they would receive course credit if they reported their behavior 10 times, or if they reported their behavior fewer than 10 times but were still in the top 50% of participants' ranking based on how many times they reported, among those who missed at least one occasion.

Results

According to our pre-registered analysis plan, we only analyzed responses from participants who visited an elevated floor in a university building equipped with both stairs and an elevator. The data were split into a training set (80%) and a test set (20%), in a way that every observation belonging to the same participant was included in only one of the sets. The training set was used for model estimation, and the test set was used to see how well the model performed on new data.

The strength of the linear relationship between the contextual factors was examined by calculating Pearson's correlation coefficients between the factors. The results showed that the highest correlation was between *Environmental consciousness* and *Health*, $r = 0.77$ (for the more detailed results, see *Table 2*).

Table 2. Correlations between Contextual Factors

Correlations of contextual factors

	Speed	*Laziness*	*Destination floor*	*Fatigue*	*Luggage*	*Elevator speed*	*Environmental consciousness*	*Temperature*	*Appeal*	*Number of people waiting for the elevator*	*Health*
Speed											
Laziness	0.24										
Destination floor	0.37	0.25									
Fatigue	0.32	0.53	0.32								
Luggage	0.25	0.29	0.18	0.35							
Elevator speed	0.09	0.05	0.25	0.04	0.01						
Environmental consciousness	-0.19	-0.18	-0.15	-0.01	0.02	0.07					
Temperature	0.13	0.17	0.04	0.25	0.37	0.07	0.17				
Appeal	-0.11	-0.04	-0.22	-0.00	0.16	-0.12	0.44	0.14			
Number of people waiting for the elevator	0.07	-0.01	0.17	0.02	0.11	0.09	-0.02	0.05	-0.02		
Health	-0.18	-0.18	-0.10	-0.01	0.04	0.08	0.81	0.15	0.45	-0.05	

Computed correlation used pearson-method with pairwise-deletion.

To examine the extent to which contextual factors influenced the choices made between stairs and elevators, we defined a mixed effect logistic regression model with the choice between stairs and elevators as the dependent variable and the measured contextual factors as independent variables. Visited buildings and IDs were treated as random effects. *Speed* and *Destination floor* were allowed to have varying slopes between different IDs, as it was plausible that these factors would have different effects on different individuals.

We applied Lasso regularization to improve the interpretability and prediction accuracy of the regression models by selecting only a subset of variables, rather than using all of them, in the final model. The lambda parameter for the Lasso regularization was chosen based on BIC values. *Temperature* and *Number of people waiting for the elevator* added the least amount of information, so their regression coefficients were penalized the most by the regularization process and were reduced to 0.

Next, we wanted to estimate how well the model explained the variation in individuals' choices. To do this, we calculated the squared correlation coefficient between the predicted values and the measured values, $R^2 = 0.76$, to estimate the variance in choosing the stairs or elevator explained by the model.

Table 2. Regression coefficients with Standard Errors, and Odds Ratios of Contextual Factors

Variables	*b*	*95% CI*		*SE*	*OR*
		lower	*upper*		
Intercept	1.57	1.23	1.91	0.17	
Peers - elevator	-2.48	-3.00	-1.96	0.27	4.80
Peers - stairs	1.35	0.71	1.98	0.32	3.85
Destination floor	-1.30	-1.54	-1.05	0.13	0.27
Environmental consciousness	1.06	0.75	1.38	0.16	2.90
Laziness	-0.89	-1.13	-0.64	0.12	0.41
Health	0.80	0.51	1.10	0.15	2.23
Elevator speed	-0.39	-0.60	-0.18	0.11	0.68
Speed	-0.31	-0.53	-0.09	0.11	0.73
Number of people waiting for the elevator	0.25	0.03	0.47	0.11	1.29

Luggage	-0.24	-0.46	-0.02	0.11	0.79
Appeal	0.22	-0.00	0.45	0.12	1.25
Fatigue	-0.03	-0.29	0.23	0.13	0.97
Temperature	0.00				1.00

Note. Variables are ordered by the absolute value of *b*, from largest to smallest. ORs > 1 indicate an increase in the odds of choosing the stairs, while ORs < 1 indicate a decrease in the odds of choosing the stairs when the given feature is increased. In this case, the influence of *Peers* is represented by two coefficients. These values indicate how the probability of choosing the stairs changed when peers opted for the elevator or stairs, compared to when there were no peers present.

Finally, we wanted to estimate the success of our model in correctly categorizing new data. We compared the model's predictions on the test data to the real decisions to assess the accuracy of the model. We used a probability threshold of .5, where predicted probabilities higher than .5 were categorized as someone choosing the stairs rather than the elevator. The results showed that the model correctly categorized 91.43% of the new cases.

Step 3 - Assessing Beliefs about the Influence of Factors

In step 3 of the Choice Context Exploration, we aimed to measure the extent to which people believed that the collected contextual factors influenced their choices between stairs and elevators, and to explore whether people could be divided into groups that shared similar beliefs about what influenced their choices.

Method

We collected data from 373 (298 female, 1 did not wish to answer) university students from the same subject pool as in Step 1 and Step 2 (M_{age} = 21.86 years, *SD* = 3.41). The research plan was approved by the local institutional ethical review board.

An online survey was created to assess beliefs about the perceived importance of the potential contextual factors defined in step 1. The first question asked whether the

participants had any physical or mental condition that would prevent them from using the stairs or the elevator (e.g., injury, claustrophobia). This accounted for two of our previously defined factors, *Physical limitations* and *Claustrophobia and technical problems*. We did not ask any further questions of participants who reported having their options limited by physical or psychological conditions, as there would have been very low variance in their answers. Each of the 13 factors defined in step 2 was assessed with a single item, measuring how important the given factor was believed to be for the participants when choosing between elevators and stairs on a Likert-type scale ranging from 0 - *Not important at all* to 10 - *Very important*. Additionally, we included an easy arithmetic task between the questions to detect and filter out inattentive responses. The survey is available at https://osf.io/y7pmd/. Participants filled out the questionnaire either before or after the behavioral measurements of Step 2.

Results

Descriptive statistics of variables measuring beliefs are presented in *Table 3*. In order to explore individual differences regarding the factors influencing people's choices, we subjected the variables measuring the beliefs of participants to model-based clustering. This method assumes that the data come from multiple distributions, and aims to find the number of these clusters by finding their means and covariance matrices. We calculated the 10 differently parameterized models available in the *mclust* package in *R* (Scrucca et al., 2016), and compared the BIC values of these models. The model with the lowest BIC value was chosen.

Table 3. Descriptive Statistics of Beliefs

Descriptive statistics - Beliefs

variable	*n*	*mean*	*sd*	*median*	*skew*	*kurtosis*	*se*
Destination floor	373	7.79	2.67	9	-1.38	1.14	0.14
Luggage	373	7.44	2.76	8	-1.16	0.65	0.14
Speed	373	7.25	2.79	8	-1.02	0.29	0.14
Fatigue	373	6.80	2.65	7	-0.81	-0.03	0.14
Number of people waiting for the elevator	373	6.56	2.80	7	-0.82	-0.08	0.14
Peers	373	5.95	3.02	7	-0.65	-0.61	0.16
Laziness	373	5.85	3.01	6	-0.40	-0.86	0.16
Elevator speed	373	4.86	2.96	5	-0.15	-1.04	0.15
Health	373	4.54	2.96	5	-0.02	-0.97	0.15
Temperature	373	4.38	3.27	4	0.06	-1.35	0.17
Appeal	373	4.28	3.03	5	0.08	-1.08	0.16
Environmental consciousness	373	3.82	2.95	3	0.42	-0.82	0.15
Distance from entrance	373	3.70	3.01	3	0.31	-1.06	0.16

The results show that participants can be divided into three groups in which its members hold similar beliefs about what influences their choices of stairs or elevators. Three clusters were defined and every cluster was named based on the pattern of factors. The first cluster, Efficiency group (N=118), had the highest mean scores on every scale except for *Health* and *Environmental consciousness*. *Speed, Luggage, Laziness, Fatigue* and *Destination floor* scales have the highest mean scores. The second cluster, the Health & Environment group (N = 68), had their highest mean scores on the *Health* and *Environmental consciousness* scales; every other mean score was low. In the third cluster, No priority group (N = 187), there was no substantial difference in the group mean scores, and between the group mean scores and the sample mean scores.

Figure 2. Believed importance mean scores of the potential influencing factors. Bars show the mean scores across subjects, while the error bars represent 95% confidence intervals.

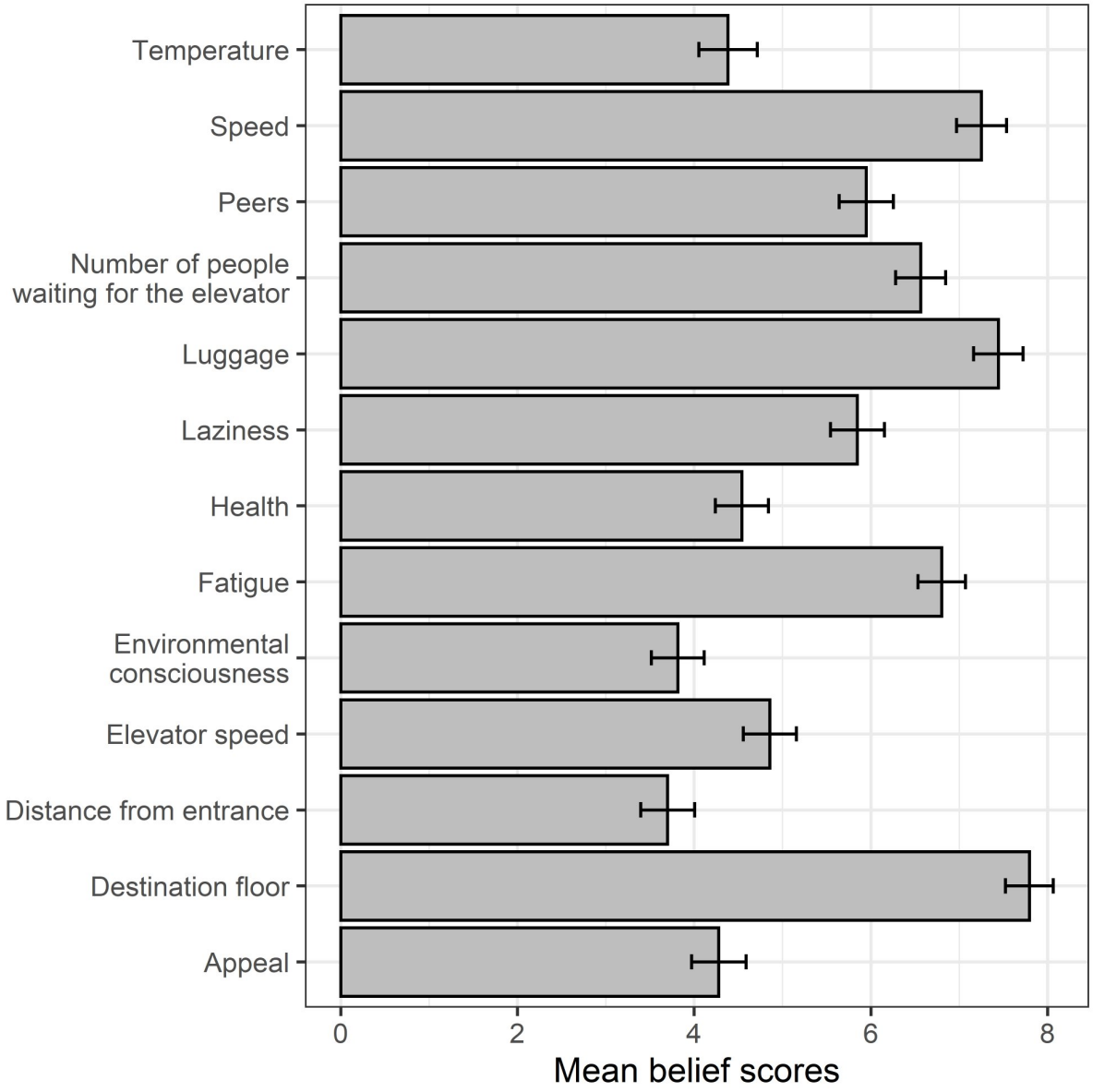
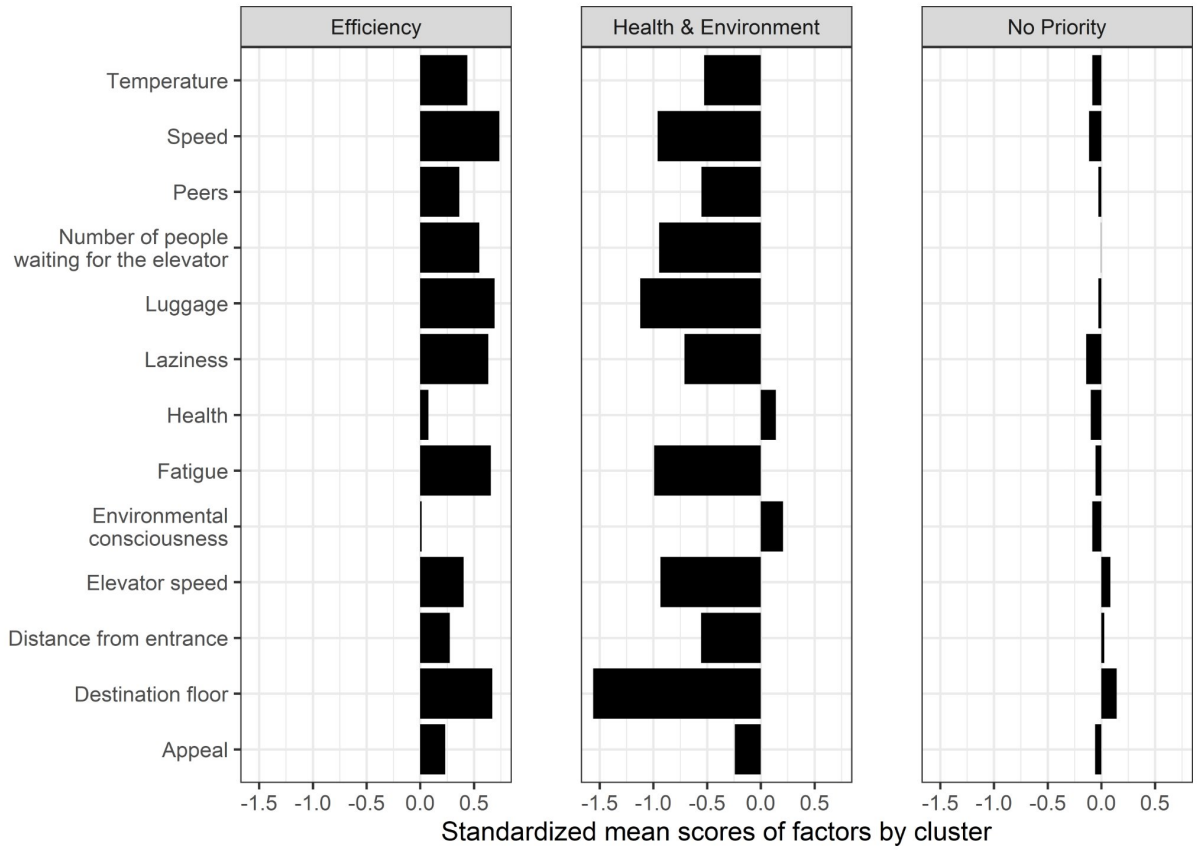


Figure 3. Believed importance standardized mean scores of the potential influencing factors by clusters. Bars show the standardized mean scores across subjects.



Comparative analysis of Step 2 and Step 3 results

We wanted to examine whether people are correct in their beliefs about which contextual factors influence them the most. To do this, we sorted the 5 factors believed to be the most influential, based on their mean scores, as well as the 5 most influential factors according to behavioral measurements, based on the *b* regression coefficients from highest to lowest. Then, we compared their rankings. The first 5 beliefs with the greatest sample means were *Destination floor*, *Luggage*, *Speed*, *Fatigue*, and *Number of people waiting for the elevator*. From the behavioral model, the variables with the greatest b coefficients were *Peers*, *Destination floor*, *Environmental consciousness*, *Health*, and *Laziness*. If we consider the exact rankings, there is no match between the two sets. If we only consider whether the given factor is in both sets, 20% of the factors are common.

Discussion

Practitioners of choice architecture interventions often face the challenge of adapting interventions to new contexts without knowing how well they will perform in those

contexts. This often leads to a trial-and-error approach, which decreases the predictability of the success of interventions. Oftentimes, the importance of context in the success of interventions is not recognized. We argue that planning interventions should start with a thorough investigation of the contextual factors of any targeted choice. This paper introduces a new procedure, the Choice Context Exploration, to help intervention researchers explore the actual and perceived contextual factors of situational choices. The three steps of the procedure have been demonstrated in a specific situation: university students' choice between using the elevator or the stairs.

In step 1, we collected 15 potential contextual factors that might influence people when choosing between stairs and elevators. In step 2, using a survey based on these factors, we estimated the effect of these factors on the participants' behavior. Based on this estimation, we identified the most influential factors regarding their contribution to the studied choice. The choices of peers, the destination floor, as well as how environmentally conscious a person is and how healthy a person aspires to be seem to have the greatest effect. The results of the analysis suggest that using the Choice Context Exploration procedure, it is possible to accurately predict, in our case over 90%, whether someone will choose the stairs or the elevator based on contextual information.

In step 3, we found that participants can be divided into three discernible groups with members who hold similar beliefs about what influences their choices between using the stairs or elevators. The "Comfort-driven" group believed that their choices are mainly based on factors such as which option they think is faster, whether they have luggage, how lazy they feel, how fatigued they are, and which floor they want to go to. The "Principles-driven" group seemed to consider which option is healthier and which is better for environmental reasons. The "No priority" group, which was the most numerous, believed that they care equally about almost every factor.

We also compared people's beliefs and behavior. People seemed to correctly assess only that the destination floor is important in their choice. However, they held false beliefs about the other influencing factors. This lack of correctly evaluated factors implies that people are not really aware of what matters most to them when deciding between using the stairs or the elevator.

What benefits did we gain from using the Choice Context Exploration in this situation? Without exploring the context of our choice, we might have missed some potentially influencing factors and would have had no way of knowing their strength. This would have left us without any guidance on which factors to target with our intervention. Additionally, if we had relied on people's apparently false beliefs without exploring them, we could have been misled about what contextual factors are important in their choice.

After using the Choice Context Exploration to gather relevant information, planning an intervention for this choice situation would be much easier: we already know the main factors that contribute to the choices made, the beliefs of the target population, and any discrepancies between the two. Based on the behavioral measurements, we can identify the factors that have the greatest effect on the target behavior, in this case the behavior of peers. It may be worth designing future interventions around this contextual factor, such as using stimuli that emphasize the importance of the decisions of peers. Understanding the beliefs of the target population can be directly applied to intervention planning. Based on our knowledge about the belief groups in our population, we may want to tailor our interventions to target one group more than the others; for example, Principle-driven people may be more influenced by interventions that build on their identities, while Comfort-driven individuals may require more costly interventions that change the environment itself to influence their choices.

Choice Context Exploration can be useful in situations where choice architects face a new target choice, a new environment, or a new population. The procedure can be particularly beneficial when the prevalent "trial and error" strategy of intervention selection would be too expensive or time-consuming. By exploring contextual influences in advance, the expenses of finding a working intervention can be reduced, as the set of implementable working interventions decreases with a better understanding of the default choice situation. It is also a beneficial option when the risk of failed and counter-productive interventions needs to be minimized to prevent negative consequences. In most cases, even if we have the resources and time to test every intervention first, repeatedly subjecting the target population to different interventions may diminish their effectiveness and make it difficult to identify the cause of an existing effect. Our main motivation behind the creation of the Choice Context Exploration is to define a set of steps that result in accurate predictions of what people might do in a given situation. One

of the main advantages of the Choice Context Exploration is that unlike in other frameworks, such as MINDSPACE (Dolan et al., 2010), EAST (Algate et al., n.d.), or BASIC (Schmidt et al., 2016), we test our models in a predictive framework and can measure the success of the collection of potential influencing factors by measuring prediction accuracy.

The Choice Context Exploration has several limitations. It was designed to provide a general overview of contextual factors in a choice situation, but new influencing factors may emerge that were previously unaccounted for, and identified factors may change their effect over time. To study these dynamic changes in influential factors, longitudinal research designs may be used. Although the Choice Context Exploration focuses on identifying influential factors, their influence may be a result of unexplored interactions. It is important to tailor models defined in the Choice Context Exploration to be able to describe the relationships between factors; in some situations, this may be achieved through the use of hierarchical models or Structural Equation Modeling, among other methods.

The sample of experts who replied to our inquiries was small, and their level of expertise may vary. Our review of the literature was not systematic. The evaluation and categorization of collected answers by only one person is suboptimal, but content analysis can be done at several levels of abstraction and there is no one true solution for a given set of answers. Our solution produced a useful set of concepts that we could use to make highly accurate predictions about behavior. Another limitation is the method of acquiring contextual factors from laypeople. Students were asked about reasons for choosing stairs or elevators in general, and they may have thought about any situation - including hotels, for example - but their behavior was analyzed in a specific context. As a result, the effects of elevator speed, for example, may not be generalizable to other, unknown environments. Our studies are based on responses from university students, and so the results of our analyses cannot be generalized to the population level. However, this was not our intention; rather, we aimed to explore and measure contextual information in a specific situation. The use of self-reported measurements might be seen as a limitation. The reason we chose to request self-reported stairs or elevator use is because it was not feasible to interview every single person, or have someone watch them at all times. Also, with

observation alone, we could not access the mental states and opinions of participants, which we see as an important aspect.

The sample sizes of our studies were based on availability, and as our studies are exploratory in nature with the purpose of informing future confirmatory research, sample size is of less concern as long as the predictive models converge and give interpretable results. In our case, we had to simplify our model in order to get a valid model estimation in Study 2, which means that our sample size was too small. In future research, it may be beneficial to designate a sample size based on model complexity. One of the main problems in the field of behavioral change interventions is overgeneralization of results and failure to account for heterogeneity. Therefore, further studies should explore the effects described in this article on different subpopulations - the method described is well suited for this task. There are other aspects that were not studied here but could serve as further points of discussion in intervention design, such as the degree of involvement of the target population, the risks presented by each choice, and whether the choice is unique or has to be made multiple times.

Our results suggest that using the Choice Context Exploration in the planning stage of future interventions could be beneficial. Our study introduced a method for investigating the contextual factors influencing a specific choice situation, using the example of choosing between stairs and elevators. We showed that by following a systematic and thorough procedure, it is possible to identify the strongest contextual factors affecting the decision to use stairs or elevators and use this information to accurately predict these choices. Despite its limitations, the proposed procedure appears to be effective in increasing our understanding of choice situations and helping us design more effective interventions. Further research should involve using the Choice Context Exploration in different environments and examining the moderating factors when implementing nudges.

References

Algate, F., Gallagher, R., Nguyen, S., Ruda, S., & Sanders, M. (n.d.). *EAST*.

Bellicha, A., Kieusseian, A., Fontvieille, A.-M., Tataranni, A., Charreire, H., & Oppert, J.-M. (2015). Stair-use interventions in worksites and public settings—A systematic review of effectiveness and external validity. *Preventive Medicine*, *70*, 3–13.

Brandon, A., Ferraro, P., List, J., Metcalfe, R., Price, M., & Rundhammer, F. (2017). *Do The Effects of Social Nudges Persist? Theory and Evidence from 38 Natural Field Experiments* (No. w23277; p. w23277). National Bureau of Economic Research. https://doi.org/10.3386/w23277

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Dolan, P., Hallsworth, M., Halpern, D., King, D., & Vlaev, I. (2010). *MINDSPACE: influencing behaviour for public policy*.

Duflo, E., Kremer, M., & Robinson, J. (2011). Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. *American Economic Review*, *101*(6), 2350–2390. https://doi.org/10.1257/aer.101.6.2350

Jennings, C. A., Yun, L., Loitz, C. C., Lee, E.-Y., & Mummery, W. K. (2017). A systematic review of interventions to increase stair use. *American Journal of Preventive Medicine*, *52*(1), 106–114.

John, P., Cotterill, S., Richardson, L., Moseley, A., Smith, G., Stoker, G., Wales, C., Liu, H., & Nomura, H. (2013). *Nudge, nudge, think, think: Experimenting with ways to change civic behaviour*. A&C Black.

Ly, K., Mažar, N., Zhao, M., & Soman, D. (2013). *Nudging*.

Marteau, T. M., Fletcher, P. C., Hollands, G. J., & Munafò, M. (2020). Changing behavior by changing environments. In *The handbook of behavior change*. Cambridge University Press.

Nilsen P., Bernhardsson, S. (2019). *Context matters in implementation science: a scoping review of determinant frameworks that describe contextual determinants for implementation outcomes.* BMC Health Serv Res. *19*(1):189.

Puchta, C., & Potter, J. (2004). *Focus group practice*. Sage.

Rogers, L., De Brún, A., & McAuliffe, E. (2020). Defining and assessing context in healthcare implementation studies: a systematic review. *BMC Health Services Research*, *20*(1), 1-24.

Schmidt, K., Schuldt-Jensen, J., Aarestrup, S. C., Jensen, A. R., Skov, K. L., & Hansen, P. G. (2016). *NUDGING SMOKE IN AIRPORTS*. 8.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289–317.

Silva, A., & John, P. (2017). Social norms don't always work: An experiment to encourage more efficient fees collection for students. *PLOS ONE*, *12*(5), e0177354. https://doi.org/10.1371/journal.pone.0177354

Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., & Aczel, B. (2018). A Systematic Scoping Review of the Choice Architecture Movement: Toward Understanding When and Why Nudges Work: Systematic Scoping Review of the Nudge Movement. *Journal of Behavioral Decision Making*, *31*(3), 355–366. https://doi.org/10.1002/bdm.2035

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Tipton, E., Yeager, D. S., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. *Experimental Methods in Survey Research: Techniques That Combine Random Sampling with Random Assignment*, 435–456.

Supplementary materials

Demographics of the Step 1 survey
In the Step 1 survey, participants (M_{age} = 22.00 years, SD = 2.32) received course credits as compensation.

Items of the Step 1 survey
Participants were asked to give as detailed answers as possible to the following questions:

"In your opinion, what are the influencing factors when people choose stairs over elevators/elevators over the stairs?"
"In your opinion, what are the influencing factors when you choose stairs over elevators/elevators over the stairs?"
“Did your stair usage change over the years?”

Answer frequencies of the Step 1 survey

Category	Number of times mentioned
Speed	316
Comfort/Laziness	312
Health/Sports	299
Destination floor	267
Fatigue	157
Physical limitations	157

Luggage	133
Claustrophobia and technical problems	132
Number of people waiting for the elevator	129
Elevator availability	81
Peers	50
Environmental consciousness	15
Elevator/stairs placement	11
Temperature	11

Experts' survey in Step 1

"In your opinion, what are the factors which can influence people's decision making (in any direction) when choosing between stairs and elevators when going up?"

Based on the answers from experts, the list was expanded by one factor, namely "Attractiveness" - the only factor mentioned by the experts, but not by the students.

Deviations from the preregistration

In the preregistration document, we specified the mixed effects logistic regression model calculated in Step 2 to allow random slopes for Peers and Elevator/Stairs placement, too, besides Speed and Destination floor. In order to simplify our model, we decided not to allow Peers and Elevator/Stairs placement to have random slopes. We also specified that if correlations between measured potential influencing factors are higher than .75, the factors in question would be merged into a principal factor, or multiple principal factors, unless it would not make sense from a theoretical point of view. We did not calculate a principal component in the single case where Pearson's r was higher than .75, between Environmental consciousness and Health, because we thought it appropriate not to merge the two variables. With these exceptions, we followed the preregistered methods.

Methods

There might be an overlap between the samples of Study 1 and Study 2, because they were recruited from the same participant pool, and we did not prohibit participants of Study 1 from applying for Study 2. Also the wording about not including the Physical limitations and Claustrophobia and technical problems might be misleading. We, in fact, did not include these factors in our first questionnaire of Study 2, rather than including them and then excluding them later.

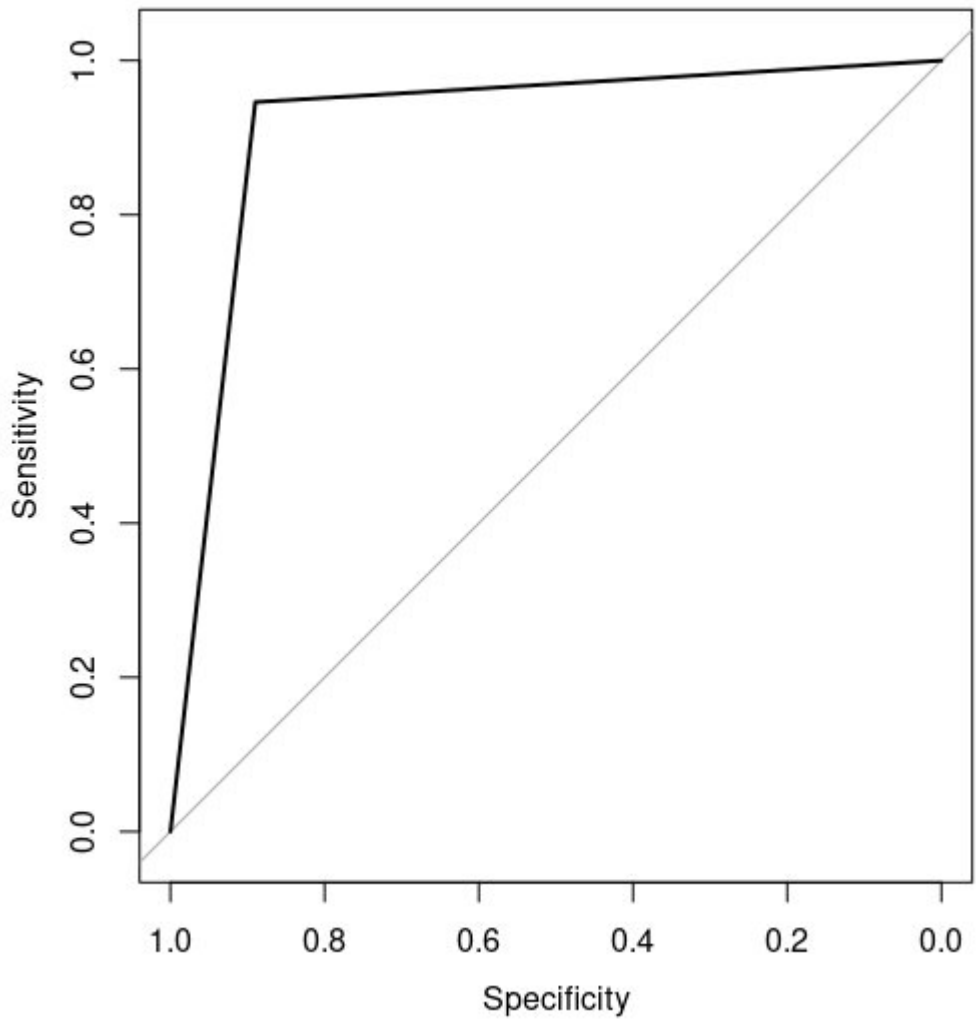
Analyses

The data was split into training and test sets only after variables have been standardized. This is suboptimal, as it could lead to information leakage. However, separating the test set in a way that answers from the same participant are not split between the training and test set is a good practice that prevents leakage from this source. Data from a specific participant were only used either in the training or the test set. Correlation coefficients of the predictors used in Study 2 were calculated using the test set only. The predictor Speed refers to whether the participant was in a hurry when they made a choice between the stairs and the elevator. The R2 of the LASSO model was computed as the squared biserial correlation coefficient between the predictions on the test set and the true choices in the test set.

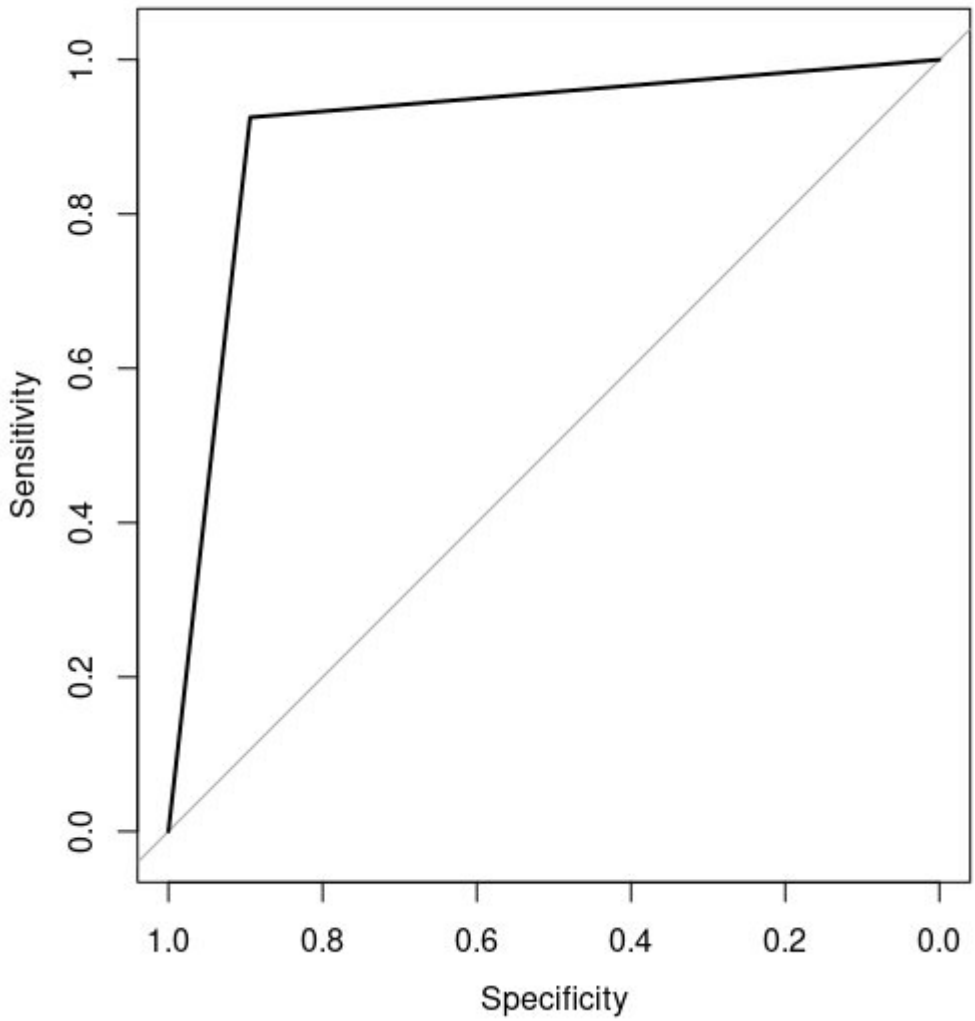
We calculated additional accuracy metrics for the predictions of the LASSO model, and we also calculated a mixed-effect glm model without LASSO regularization, for comparison. We report these metrics for both the training and test sets.

metric	LASSO - training	LASSO - test	Mixed GLM - training	Mixed GLM - test
Base rate	0.6105	0.6505	0.6105	0.6505
Accuracy	0.9242	0.9144	0.9653	0.9120
Kappa	0.8399	0.8131	0.9270	0.8066

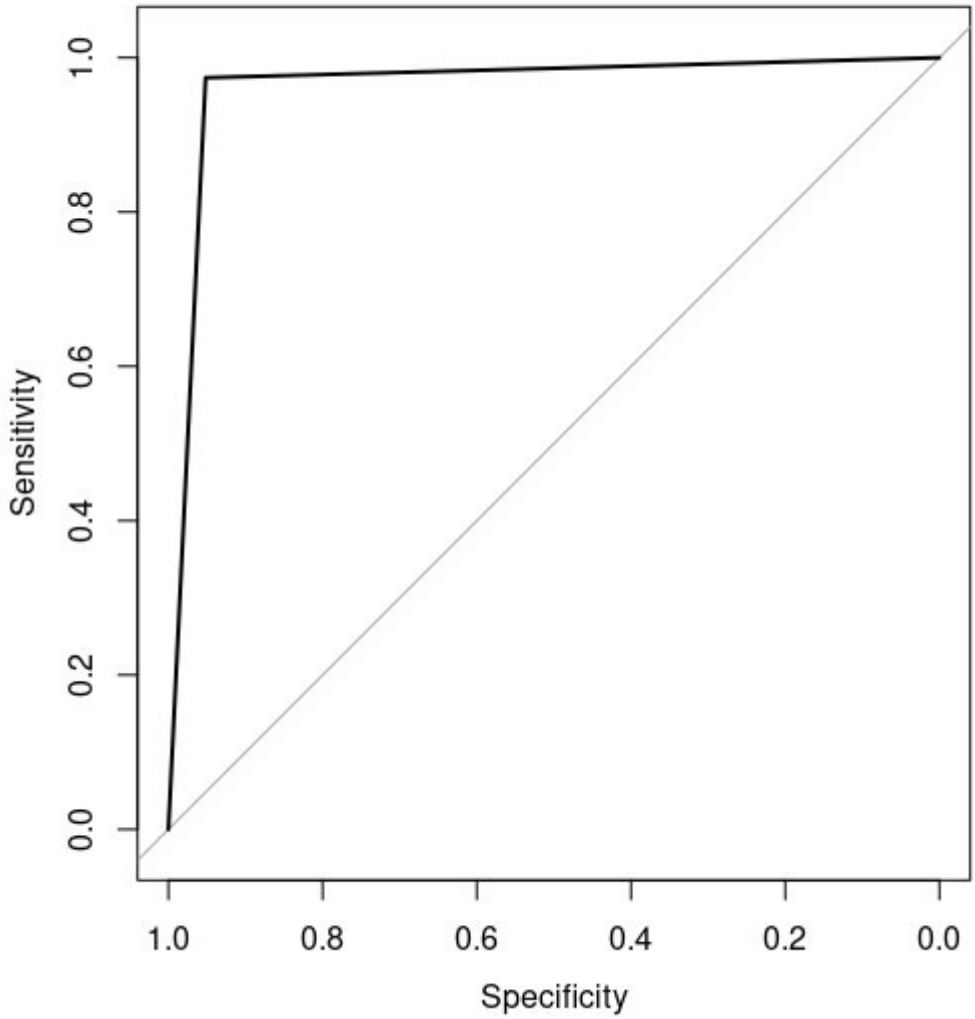
Sensitivity	0.8899	0.8940	0.9517	0.8742
Specificity	0.9461	0.9253	0.9740	0.9324
Pos predictive value	0.9133	0.8654	0.9590	0.8742
Negative predictive value	0.9309	0.9420	0.9693	0.9324
F1	0.9015	0.8795	0.9553	0.8742
Balanced accuracy	0.9180	0.9097	0.963	0.9033



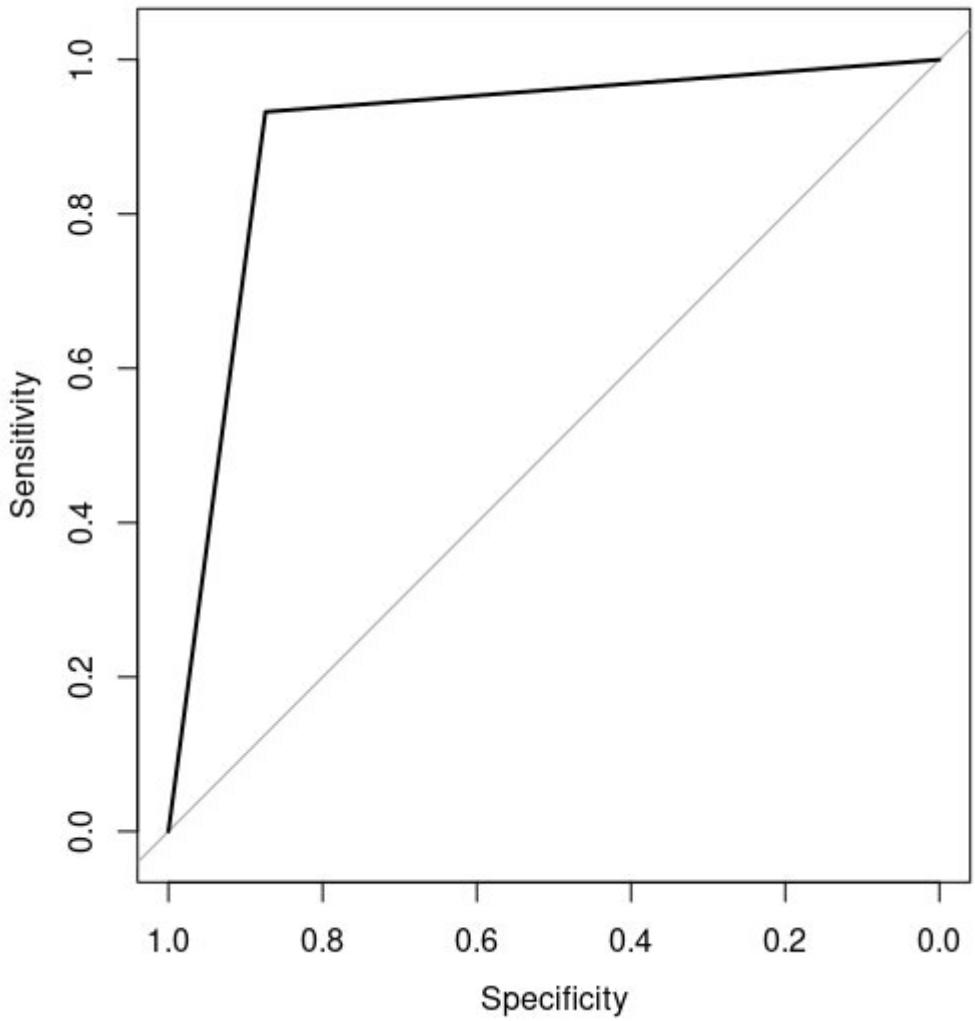
ROC curve of the LASSO model on the training set.



ROC curve of the LASSO model on the test set.

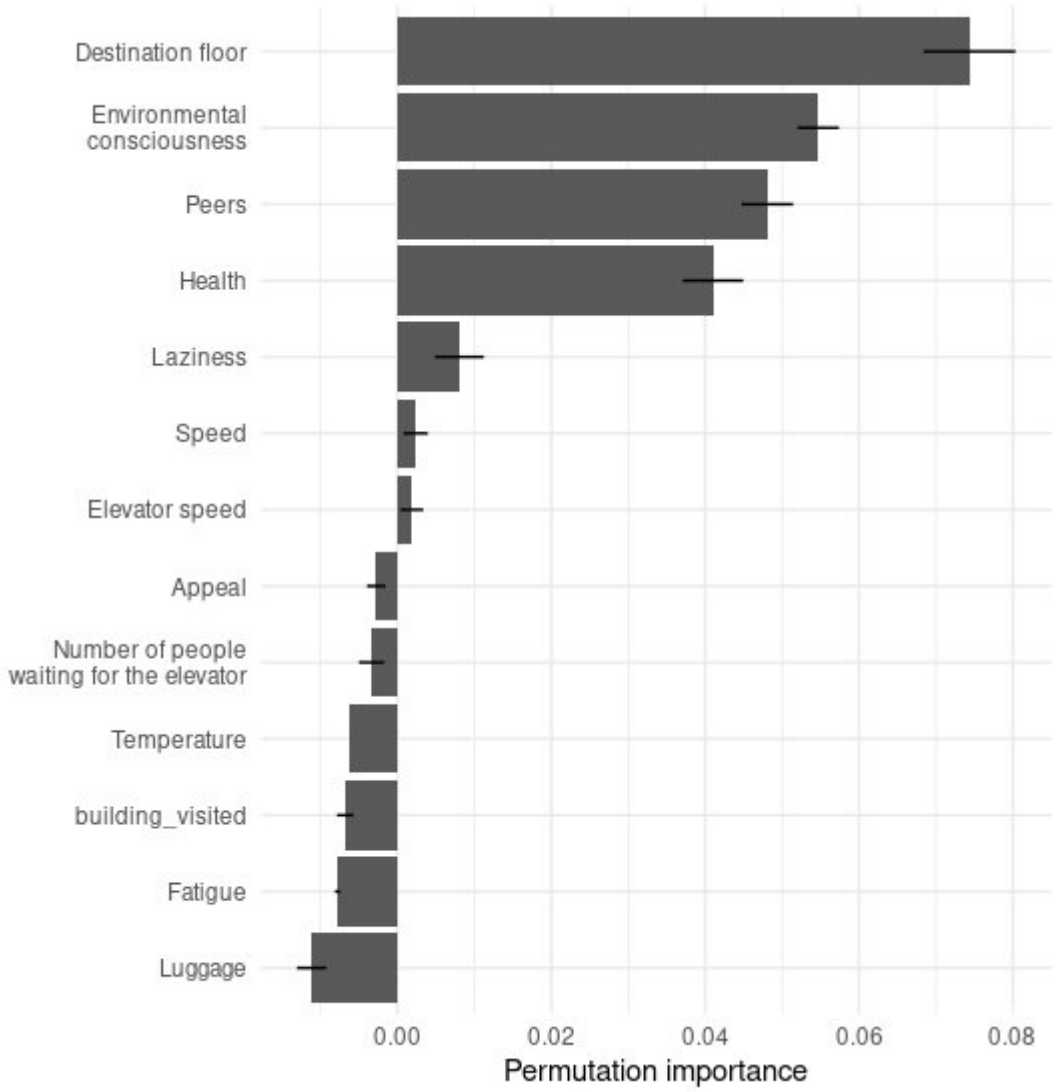


ROC curve of the mixed-effect GLM model on the training set.



ROC curve of the mixed-effect GLM model on the test set.

We also calculated permutation importance scores for the predictors. These are better suited to show which predictor was important, compared to ranking regression coefficients - also because one of the predictors was categorical, and comparing coefficients as they are might be misleading.



Permutation importance scores of predictors. The scores are the average decrease in the ROC AUC of 10 permutations, when the given variable was shuffled.

In order to be able to compare the differences of the predictive powers of reported behavior and beliefs, we defined two models to predict reported choice between stairs and elevator: one has the top 5 variables based on permutation importance: Destination floor, Environmental consciousness, Peers, Health, and Laziness, while the other has the top 5 based on mean belief scores: Luggage, Destination floor, Speed, Fatigue, and the Number of people waiting for the elevator.

Metric	Behavioral train	Behavioral test	Belief train	Belief test
Base rate	0.6105	0.6505	0.6105	0.6505

Accuracy	0.9636	0.9005	0.9142	0.7847
Kappa	0.9232	0.7801	0.8184	0.5494
Sensitivity	0.9472	0.8477	0.8718	0.8013
Specificity	0.9740	0.9288	0.9413	0.7758
Positive predictive value	0.9588	0.8649	0.9045	0.6576
Negative predictive value	0.9666	0.9190	0.9200	0.8790
F1	0.9530	0.8562	0.8879	0.7224
Balanced accuracy	0.9606	0.8883	0.9065	0.7886

The model based on beliefs has a 78.9% balanced accuracy, while the one based on reported behavior has 88.8%. Also, both models predicted elevator use more accurately. The model trained on belief data cannot predict better than the base rate, while the reported behavior-based model can.

Discussion

An important limitation of Steps 2 and 3 was that they used self-reported data. People could choose to respond or behave consistently for multiple reasons that would deviate from their normal behavior. In Step 3, we incorrectly stated that people in the "No priority" group don't have a clear priority across the factors. In fact they might well have a priority, which might even have shown up in their responses but maybe the clustering model was unable to classify them to the other two groups.

Regarding the discrepancy between reported behavior and beliefs, based on new analyses, we see a difference in the predictive power of the two. Notably, the model based on beliefs predicted staircase use with the same accuracy as the base rate.

The biggest limitation of our research presented here is that we conducted a repated measures post-event inquiry about potential influences that might have been influenced by the very fact of behavior. Because of this, the conclusions about the possible prediction of behavior are overstated. In our opinion, the results of this research are better framed as

a successful attempt in identifying strong relationships between the retrospectively reported behavior and reported context.

Chapter II.

Hajdu, N., Schmidt, K., Acs, G., Röer, J. P., Mirisola, A., Giammusso, I., ... & Szaszi, B. (2022). Contextual factors predicting compliance behavior during the COVID-19 pandemic: A machine learning analysis on survey data from 16 countries. *Plos one*, *17*(11), e0276970.

Nandor Hajdu[1,2,*], Kathleen Schmidt[3], Gergely Acs[4], Jan P. Röer[5], Alberto Mirisola[6], Isabella Giammusso[6], Patrícia Arriaga[7], Rafael Ribeiro[7], Dmitrii Dubrov[8], Dmitry Grigoryev[8], Nwadiogo C. Arinze[9], Martin Voracek[10], Stefan Stieger[11], Matus Adamkovic[12,13], Mahmoud Elsherif[14], Bettina M. J. Kern[10,15], Krystian Barzykowski[16], Ewa Ilczuk[16], Marcel Martončik[12], Ivan Ropovik[17,18], Susana Ruiz-Fernandez [19,20], Gabriel Baník[12], José Luis Ulloa[21], Balazs Aczel[2¶], Barnabas Szaszi[2¶]

[1]Doctoral School of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

[2]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

[3]Ashland University, Ashland, OH, United States of America

[4]Department of Networked Systems and Services, Budapest University of Technology and Economics, Budapest, Hungary

[5]Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

[6]Department of Psychology, Educational Science and Human Movement, University of Palermo, Italy

[7]ISCTE-University Institute of Lisbon, CIS-IUL, Portugal

[8]National Research University Higher School of Economics, Russian Federation

[9]Alex Ekwueme Federal University, Ndufu-Alike, Nigeria

[10]Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Austria

[11]Department of Psychology and Psychodynamics, Division Psychological Methodology, Karl Landsteiner University of Health Sciences, Krems an der Donau, Austria

[12]Institute of Psychology, Faculty of Arts, University of Presov, Prešov, Slovakia

[13]Institute of Social Sciences, CSPS Slovak Academy of Sciences

[14]Department of Psychology, University of Birmingham, Birmingham, United Kingdom

[16]Department of European and Comparative Literature and Language Studies, Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria

[17]Institute of Psychology, Faculty of Philosophy, Jagiellonian University, Krakow, Poland

[18]Faculty of Education, Charles University, Prague, Czech Republic

[19]Faculty of Education, University of Presov, Prešov, Slovakia

[20]FOM University of Applied Sciences, Essen, Germany

[21]Leibniz-Institut für Wissensmedien, Tübingen, Germany

[22]Programa de Investigación Asociativa (PIA) en Ciencias Cognitivas, Centro de Investigación en Ciencias Cognitivas (CICC), Facultad de Psicología, Universidad de Talca, Chile.

* Corresponding author Email: hajdu.nandor93@gmail.com (NH)

¶ BA and BS are Joint Senior Authors.

Abstract

Voluntary isolation is one of the most effective methods for individuals to help prevent the transmission of diseases such as COVID-19. Understanding why people leave their homes when advised not to do so and identifying what contextual factors predict this non-compliant behavior is essential for policymakers and public health officials. To provide insight on these factors, we collected data from 42,283 individuals across 16 countries. Participants responded to items inquiring about their socio-cultural environment, such as the adherence of fellow citizens, as well as their mental states, such as their level of loneliness and boredom. We trained random forest models to predict whether someone had left their home during a one week period during which they were asked to voluntarily isolate themselves. The analyses indicated that overall, an increase in the feeling of being caged leads to an increased probability of leaving home. In addition, an increased feeling of responsibility and an increased fear of getting infected decreased the probability of leaving home. The models predicted compliance behavior with between 62% and 87% accuracy within each country's sample. In addition, we modeled factors leading to risky behavior in the pandemic context. We observed an increased probability of visiting risky places as both the anticipated number of people and the importance of the activity increased. Conversely, the probability of visiting risky places increased as the perceived putative effectiveness of social distancing *decreased*. The variance explained in our models predicting risk ranged from < .01 to .54 by county. Together, our findings can inform behavioral interventions to increase adherence to lockdown recommendations in pandemic conditions.

Keywords: COVID-19; lockdown; machine learning; multi-national study, random forests; social distancing; voluntary isolation

Introduction

When no treatment or vaccine is available to prevent transmission, behavioral measures may be the most effective means of containing a disease ((1);, (2)). One such approach is to ensure that people minimize contact with other individuals, either by keeping a safe distance from other people in public places or by staying at home. However, maintaining sufficient compliance with these rules and regulations is difficult, especially for extended periods of time (3). In order to counter the spread of a disease, understanding which factors influence people's compliance with confinement recommendations is essential.

Among all the different factors that could affect staying at home during a pandemic (e.g. personality traits), contextual factors are the focus of the present paper. We define contextual factors as the physical and sociocultural environment along with the intrapersonal circumstances, such as mental states, present at the time of the choice that may affect decisions. Contextual factors can accurately predict decisions in simple situations where most of this contextual information can be identified (4). Identifying the contextual factors of non-adherence to lockdown recommendations and exploring their relative predictive strength will provide insight into decisions that put individuals and communities at risk. These insights can help public health officials and policy makers design interventions to target the factors that have the largest effect on decision making. Although not labeled as such in previous research, many factors that fit our definition of contextual factors (e.g. confidence in the government to tackle the pandemic (5)) have already been studied. However, the literature is limited regarding the systematic investigation of the contextual factors that influence people's decisions to comply with confinement regulations.

Most lockdown regulations during the Covid-19 pandemic have allowed individuals to leave their residences for essential reasons. The definitions of what constitutes an essential or non-essential activity likely varies according to region, but most regulations or recommendations classify going to work (when working from home is not possible), attending school or another educational institution, shopping for groceries and medicine, seeking medical care, and exercise as essential activities that justify venturing outside (e.g., (6)). Outings for any other reason are considered to be non-essential activities (e.g. social gatherings). Here, we consider leaving home for non-essential reasons as non-compliant behavior during lockdown.

Mental states and beliefs as context

Some of the main factors that motivate individuals to leave their home during confinement are feelings of loneliness (7) and other unpleasant mental states. Boredom is also a prevalent state during social isolation and boredom proneness is a critical risk factor for non-compliance with social-distancing protocols (8). Further, adverse reactions to recommendations or requirements to stay inside may lead to feelings of captivity. This sentiment is well reflected in the oft-used metaphor of "being imprisoned" when people describe their situation during quarantine (9). These mental states likely decrease adherence to social isolation recommendations during lockdown.

General compliance with isolation rules or recommendations also appears to be influenced by attitudes and beliefs, such as thinking that taking health precautions is effective against the infection (10). Among these beliefs, perceived vulnerability, beliefs that getting COVID-19 would be disruptive, and government trust each have very small positive effects on general compliance (10). However, other factors such as trust in

policies seem to have stronger effects. Researchers have found increased mobility reduction - thus, compliance with quarantine regulations - in European regions where the levels of trust in policymakers prior to the COVID-19 pandemic was high (11). In a study exploring the effects of self-perceived risk of contracting COVID-19, fear of the virus, moral foundations, and political orientation on compliance with public health recommendations, only fear emerged as a predictor of compliance (12). The perceived infectiousness of COVID-19 may also have an effect on rule compliance; the more contagious people think COVID-19 is, the less willing they are to take social distancing measures. This counterintuitive relationship has been described as the "fatalism effect" (13). Finally, the sense of duty and responsibility could also contribute to staying at home (14) because leaving the house would be perceived as irresponsible.

Motivation to remain at home during requested social isolation periods can stem from trusting in someone or something the pandemic. People might not leave their homes because they trust the regulations to be effective or place their trust in a higher power (15). Also, generalized social trust appears to moderate the indirect effect of personality traits on rule-respecting behaviors; individuals who trust others demonstrate more compliance than those who do not (16). Expert opinion may also motivate compliance; providing people with expert information about the spreading of the virus partially corrects their misconceptions about transmission (13). Compliant people seem to perceive protective measures as effective, while non-compliant people perceive them as problematic(17). Altogether, several factors have emerged as potential predictors of non-compliant behavior in the context of the current pandemic. However, these factors have not been examined systematically across cultures.

The present research was designed to extend the literature on lockdown regulations by systematically investigating the contextual factors that influence compliance in confinement situations across cultures. First, we conducted a pilot study to identify potential contextual factors that might affect compliance with confinement recommendations. Then, in our main study, we explored how these factors influenced the behavior of participants from 16 countries using a machine learning approach. Specifically, we tested the extent to which these factors predict *(a)* compliance with confinement recommendations and *(b)* the risk-taking behaviors of non-compliant individuals.

Pilot Study

The main goal of the pilot study was to identify potential influencing factors that might have an effect on whether or not someone stays at home during a pandemic. A brief survey was used to collect qualitative data to achieve this goal.

Methods

The survey respondents were recruited from a university participant pool in Hungary that consisted of undergraduate and graduate students who received course credit as compensation. The survey was conducted in March 2020, three weeks after the lockdown measures were first locally imposed. Participants responded to open-ended questions about what influences their decisions and other peoples' when they choose to leave their home and go to a place where they might be in close physical proximity to others. To process the answers, we used inductive coding to compare responses to factors already derived from the existing literature or generated by brainstorming. For each answer, the first author decided whether the given answer contained a new type of factor. If a newly

processed answer could not be labeled as belonging to any of the registered categories, a new category was created.

Results

A total of 532 participants completed the survey. After processing all the responses, we added 1 additional factor that may influence adherence to confinement recommendations, for a total of 23: *being afraid of getting infected; feeling that staying home is the responsible behavior; feeling caged; being afraid of the consequences of getting infected; being afraid of infecting someone else; thinking that they are already a vector; feeling lonely; thinking that the pandemic will have serious economic consequences; belief in the effectiveness of social distancing; being in contact with elderly/someone with chronic illness; country leaders' communication; trust in a higher power; trust in experts' opinion; trust in people who attend the out-of-home activity; knowing people who attend the out-of-home activity; event importance; peers' opinion; family opinion; number of people attending the out-of-home activity; possibly meeting many people while getting to the site of the out-of-home activity; out-of-home activity site size; event is indoors or outdoors;* and *level of hygiene at the location of the out-of-home activity*.

Main Study

The goal of our main study was to explore the extent to which the factors identified in the pilot study predict compliance with lockdown recommendations. Also, we investigated whether the riskiness of an out-of-home activity can be predicted from contextual factors, such as the spaciousness of the place or other circumstances.

Methods

The methods and analyses for the main study were pre-registered and can be found at https://osf.io/7nfu8. Deviations from the pre-registration are detailed in the *Supporting information* section. The research plan was approved by the lead authors' local institutional ethical review board. The data were collected between April 29, 2020 and November 12, 2020.

Participants

Participants were recruited with the collaboration of 16 research labs. Each research lab organized individual campaigns of participant recruitment through various media outlets, university participant pools, or paid participant pools. Details of recruitment methods for each lab can be found in the *Supporting information* section. In total, we recruited 43,123 participants from 102 countries; however, we only analyzed data from the 16 countries with more than 100 respondents (n = 42,283) to allow for more complex analyses, as well as more robust results from these. The countries included in the study were: Austria, Germany, Greece, Hungary, Italy, Japan, the Netherlands, Nigeria, Poland, Portugal, Romania, Russia, Slovakia, Switzerland, the UK, and the USA.

Materials and Procedures

The study was conducted online via Qualtrics. First, respondents reported their age, gender, years of education, country of residence, monthly income, and the number of people in their household. Then, participants were asked if they had left their home in the previous 7 days of the lockdown for non-essential reasons. There was a slight difference in wording between countries where there was a lockdown at the time of response and where the lockdown had already ended. In cases where there was a lockdown at the time of response, the question was: *"Did you leave your home in the last 7 days for non-essential reasons?*." Where the country did not have any restrictions in effect at the time

of the survey, the question was the following: *"Did you leave your home in the last 7 days of the lockdown for non-essential reasons? Lockdown is the period when residents in your region were asked not to leave their homes for non-essential reasons."* Participants were informed that essential reasons included: buying groceries or medicine, going to work, and seeking medical attention in case of serious illness/injury. Next, participants were asked to indicate the degree to which the statements - corresponding to each of the 24 factors identified in the pilot study - applied to them or to their activity on a 7-point Likert-type scale (1= *did not apply at all*; 7 = *completely applied*).

Event-specific items that referred to factors concerning the context of the out-of-home activity only appeared for participants who actually left their home during the investigated period. For these event-specific items, participants were asked to respond to statements about their most recent non-essential out-of-home activity. The 9 event-specific items measured were: *peer pressure to take part in the activity; the number of people present; degree of acquaintance; trust in the people present; preconception about how many people they would meet; location size; location indoors or outdoors; hygiene of the location;* and *importance of the activity.*

Event-general items (i.e., those not specific to an out-of-home activity) were shown to every respondent, regardless of whether they left their homes in the previous 7 days. For these items, participants were asked to indicate their degree of agreement with 16 statements describing *the fear of getting infected; thought that already contacted the virus, boredom, loneliness, coping with being indoors, thoughts about symptom seriousness if infected, economic consequences, putative effectiveness of social distancing, trust in a higher power, contact with elderly or someone with chronic illness,*

fear of infecting someone else, feeling of responsibility, encouragement of country leaders, encouragement of experts, adherence of fellow citizens, being up-to-date about the virus. Note that the "*contact with elderly or someone with chronic illness"* and the *"being up-to-date about the virus"* items were excluded from analyses by the lead team because they were judged not to measure context. Among these items, participants also responded to an attention-check item: *"I went to the Moon twice."*

The original English language questionnaire was translated to eleven languages by native speakers from the participating research labs. The full survey for each language is available at https://osf.io/u38zh/.

Data Analysis

To answer the question of why people leave their homes during a pandemic lockdown, we opted to use random forest models, a machine learning method (18). Random forests are popular prediction algorithms for several reasons: they are robust to the non-linearity of data, they do not require data to be normalized, and they typically provide superior prediction accuracy while mitigating overfitting without extensive parameter tuning. It is a standard method of machine learning and is frequently employed when the number of variables to consider is relatively low. However, this method has some limitations. The results are not as easy to interpret because decision trees are stochastic, which means that they can change with different runs. Random forests are made of decision trees. Each decision tree in the forest is a set of internal nodes and leaves. In the internal node, a feature is selected along which the data is split into two groups. Then, each group is subdivided iteratively, following the same rule until some condition is met on the size of the tree or the number of data points in the node. For classification problems, the criterion to select a feature can be Gini impurity or information gain. We used information gain in

our calculations. The average information gain increase is collected for each feature selected for the splits. The average of this increase over all trees in the forest is the measure of variable importance. Because a random subset of features is used for a tree, the result is also random. However, if we have many trees, then the resulting importance values should be similar to one another. We analyzed data from each country separately.

To explore the factors that predict non-compliance (i.e., leaving home for non-essential reasons), we created random forest models using the event-general items. Data were split into training and test sets in an 80-20 ratio. On our training dataset, the number of variables in each division of a tree node was between 2 and 10 and were tuned separately for every country via 10-fold cross-validation. Then, we tested how well each model performed on the test data by calculating classification accuracies. We also calculated variance importance metrics for each model. These metrics inform us of the degree of importance of a variable to predict outcomes. We used the variance importance scores based on the mean decrease in accuracy when the given variable is removed from the model.

To analyze the riskiness of activities, we first defined a "risk" score as the sum in the levels of crowdedness, size, level of hygiene, and whether the event was indoor. The greater this score the higher the risk of the activity. Next, we created random forest regression models on data from individuals who indicated that they left their homes during the lockdowns. Consequently, we could include both event-specific and event-general items in this analysis. We use the risk score as the dependent variable to estimate the influence of a factor in the decision to participate in an activity despite it being risky. Variable importance was calculated the same way as in the case of non-compliance

prediction. The greater the importance of the predictor, the more influence it has on the decisions of people to go outside despite being in a risky situation. As the dependent variable was continuous, we calculated the Root Mean Squared Errors to assess the model accuracy, and chose the model with the lowest error during hyperparameter tuning.

Results

Data of respondents who did not finish the questionnaire were excluded from the analysis (N = 13,653), along with those who failed the attention check (N = 2,387). We only considered countries with more than 100 respondents in our analysis. We also excluded those who reported the top 0.1% income in each country (N = 56), because the values were unrealistically high. After exclusions, the final sample used in the analysis included 42,283 people from 16 countries (M_{age} = 40.92 years, SD_{age} = 12.06, 50.86% female). Table 1 shows the basic descriptive information for each analyzed country.

Table 1. Sample Descriptive Statistics by Country. Left home proportion represents the proportion of people who left their homes for non-essential reasons out of all respondents.

Country	*N*	Female proportion	Left home proportion	Median income per month (USD)	Median age (years)	Median years of education
Austria	1140	0.67	0.42	2739.00	28	17
Germany	2217	0.67	0.47	3834.60	27	16
Greece	135	0.75	0.59	1314.72	50	16
Hungary	35097	0.48	0.52	1987.34	42	17

Italy	477	0.70	0.18	657.36	28	17
Japan	280	0.45	0.30	1885.92	45	16
Netherlands	117	0.56	0.64	4930.20	35	17
Nigeria	186	0.52	0.43	94.19	26	14
Poland	377	0.70	0.46	482.32	23	16
Portugal	381	0.65	0.46	2191.20	33	16
Romania	115	0.50	0.38	1591.52	41	17
Russian Federation	377	0.39	0.40	649.00	30	15
Slovakia	350	0.85	0.36	1643.40	21	15
Switzerland	151	0.52	0.62	10358.40	40	18
United Kingdom	459	0.48	0.39	4449.78	38	17
USA	424	0.46	0.51	9500.00	36.5	16

Factors predicting non-compliance

A heatmap showing the differences in relative importance for each item and country is shown in Fig 1. As shown, *Feeling of responsibility* was in the top three most important factors in 13 out of 16 countries, suggesting that it is one of the most important factors overall in predicting home confinement. The *feeling of being caged while at home* and *fear of getting infected* also had a great impact on staying at home, as they were in the top three most important factors in 10 and 11 countries, respectively.

Fig 1. Variable importance values when predicting leaving home in each country.

We calculated the permutation importance of a variable, i.e., the decrease in prediction accuracy when the given variable is randomized, while other variables are left intact. This randomization was conducted 100 times, and the average importance is reported. To provide a visual representation of the differences between the importance values of variables, we rescaled the variable importance values per country to values between 0 (least important) and 100 (most important). The darkness is based on the rescaled

importance score, grouped by country: the higher the permutation importance score of a variable in a given country, the darker the color.

Our models, based on event-general factors, were successful in predicting whether someone left their home during lockdown. Predictions were the most accurate on Italian data, where 87% of the test cases were classified correctly. The least accurate predictions were made on Portuguese data, with only a 62% accuracy. All the model accuracies are reported in Table 2.

Table 2.

Prediction Accuracies of Random Forest Models by Country. Left home - accuracy represents the percentage of correct classifications on the test set when predicting whether someone left their home. *Risk - Root Mean Squared Error* indicates the accuracy of predictions on the test set when predicting riskiness of the activity when someone left their home, while *risk - R^2* represents the proportion of variance explained by the model.

Country	left home - accuracy	risk - Root Mean Squared Error	risk - R^2
Austria	0.83	0.63	0.51
Germany	0.81	0.78	0.54
Greece	0.63	1.18	< 0.01
Hungary	0.71	0.86	0.20
Italy	0.87	2.03	0.08
Japan	0.64	1.06	0.23
Netherlands	0.70	0.91	0.40

Nigeria	0.70	1.17	< 0.01
Poland	0.72	1.22	0.20
Portugal	0.62	1.05	0.006
Romania	0.65	1.10	0.27
Russian Federation	0.74	1.21	0.04
Slovakia	0.67	1.16	0.28
Switzerland	0.67	0.78	0.09
United Kingdom	0.70	0.91	0.01
United States	0.65	0.88	0.23

We created partial dependence plots to examine whether a factor was associated with an increased or decreased probability of leaving home (Fig 2). The plots suggest that the general patterns of the results were similar between countries. Inspecting the plots of the top 3 most important variables revealed that scores on the *Feeling of responsibility* scale are negatively related to the probability of non-adherence; *Fear of getting infected* seems to decrease the probability of leaving one's home, while *Feeling caged while at home* increases the probability of leaving one's home.

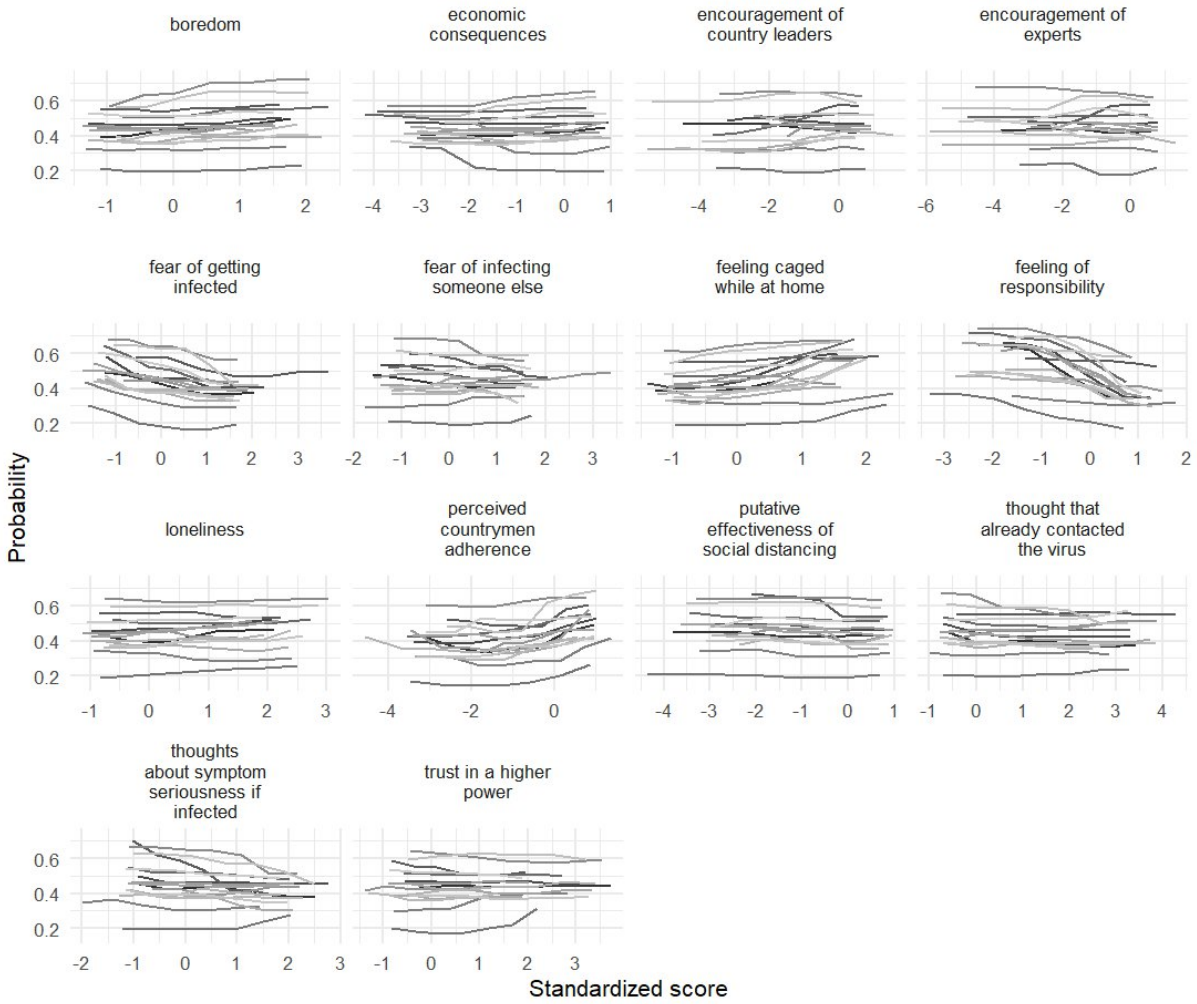


Fig 2. Partial dependence plots of variables used in the prediction of leaving home for all countries. Each line represents a different country.

We calculated how the overall prediction changes at different values of a variable by substituting real data with the same value for every participant and then calculating the mean of these predictions. This method is appropriate because the variables are uncorrelated. As a result, these predictions for different plugged-in values can be represented on a graph to see how the predictions change from one value of the independent variable to the next. Lines on Fig 2. show the average predicted probability of leaving home associated with a given value of the contextual factor in each country.

Factors predicting participation in risky activities

After analyzing the factors involved in leaving home during the lockdown, we set out to investigate the factors associated with participation in risky activities. We report the root mean squared errors and R^2 values of the final models in Table 2. Variance importance metrics were calculated for each model. A heatmap of variable importance among countries is presented in Fig 3. The results suggest that the *putative effectiveness of social distancing, activity importance, and anticipated number of people met while traveling* are the most important factors when predicting the participation in risky activities. These variables are in the top three most important variables in 5, 6, and 10 countries, respectively.

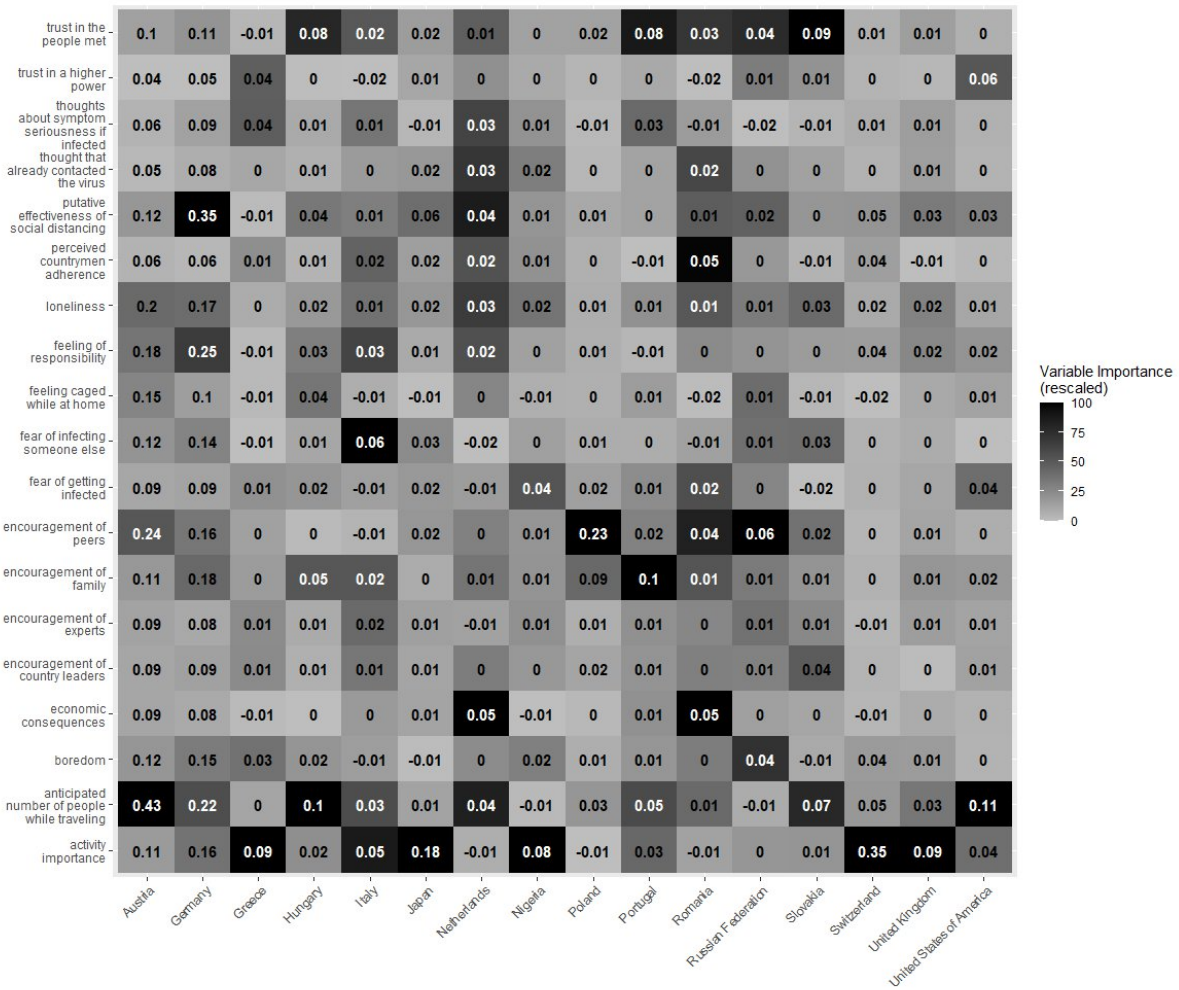


Fig 3. **Variable importances when predicting risk level of out-of-home activity in each country. The color of each cell is based on variable importance rescaled to the 0-100 range, while numbers in cells represent the original variable importance.**

Similar to Fig 1, Fig 4 shows the partial dependence plots displaying the level of riskiness associated with each factor and the change in the predicted risk score when a given variable was altered, for each country separately. The plots suggest that the general pattern of the results was similar among countries, and that, in most cases, a change in any one variable amounted to very little change in predicted risk.

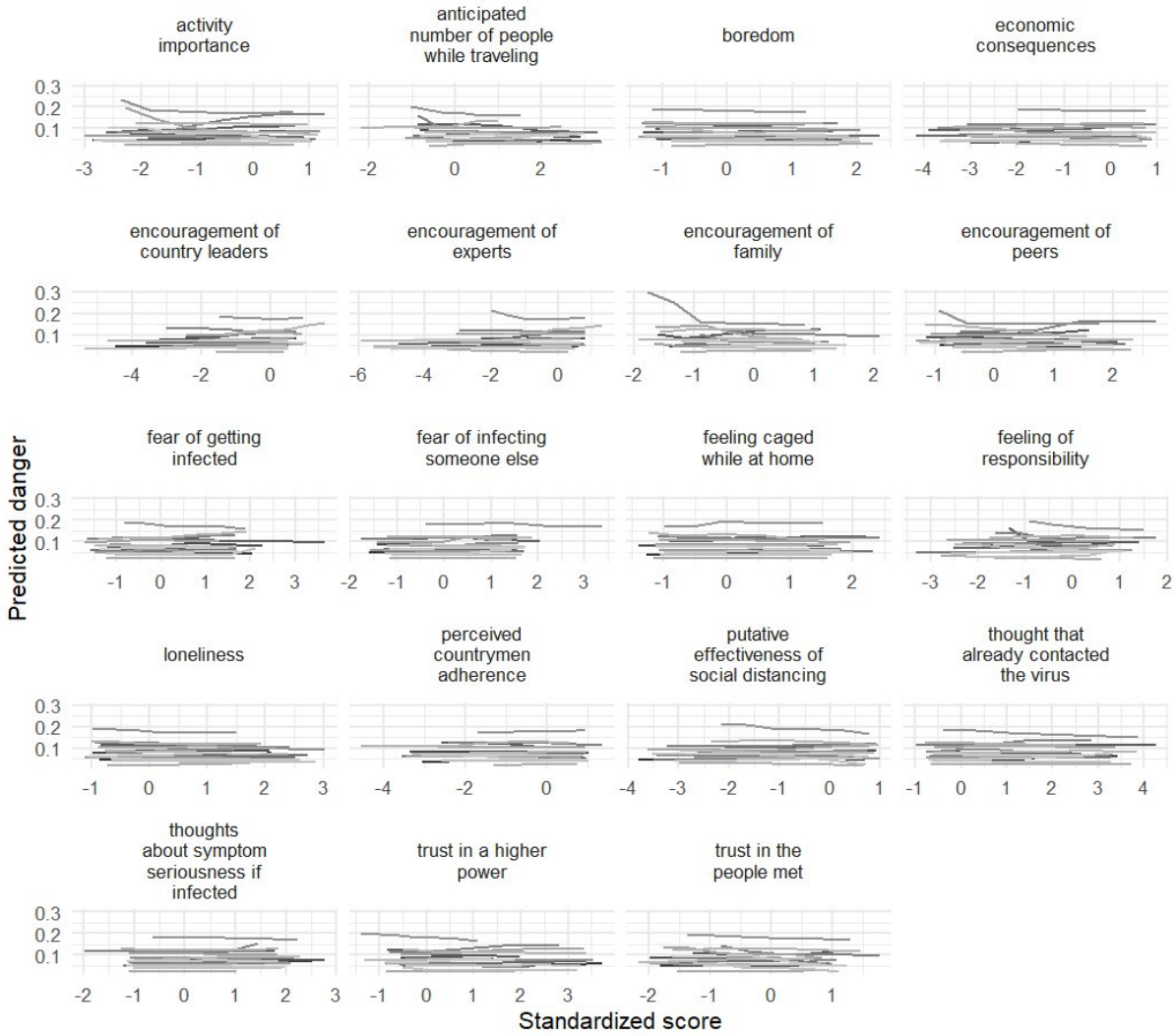


Fig 4. Partial dependence plots of variables used in the prediction of risk scores for all countries.

Discussion

The research presented here explored the importance of contextual factors in predicting decisions to stay at home during pandemic lockdowns. The factors we measured appeared to either increase or decrease the probability of leaving home across samples. In fact, the observed variables showed a consistent pattern of prediction across the 16 investigated countries, suggesting that our findings are robust and may be generalizable across cultures. *Boredom* and the *adherence of fellow citizens to regulations* increased the probability of leaving home in every country, while the *fear of getting infected* and the *feeling of responsibility* decreased the probability of leaving home in every country.

Although the examined countries differed in which factors were most important in predicting compliance with stay at home orders, some factors emerged as highly important in most of our samples. *Feeling of responsibility* was one of the top three most important factors for 13 countries. This finding suggests that feelings of obligation toward society in preventing the spread of disease increased adherence to confinement recommendations. Relative to other factors, responsibility seemed to have the largest predictive importance: when people feel responsible, they tend to stay home. However, *responsibility* had a strikingly small relationship with adherence in Japan, Switzerland, and Greece. While the *responsibility* factor was not important for these countries, the *perceived countrymen adherence* factor mattered to a great degree. Importantly, prediction accuracy was quite low for these three countries compared to the other countries overall. Perhaps other factors not explored in our study better explain compliance in these nations. The *fear of getting infected* was among the top 3 most important factors in 11 countries, but its predictive effects on leaving home were particularly accentuated in Hungary, Japan, the Netherlands, Romania, Switzerland, and

the UK and comparatively minimal in Greece, Nigeria, and the Russian Federation. Although the *feeling of being caged while at home* was among the top three most important factors in 10 countries, its effects were particularly important in the UK and Slovakia and unimportant in Japan and Greece.

Previous research has demonstrated that mental states, such as the feeling of loneliness (7) and boredom (8) are predictors of non-compliant behavior. Fear of the virus (12) has also been linked to increased compliance, along with the feeling of responsibility (14). Our study confirmed these effects and showed they are similar across countries.

Our analyses of the factors predicting activity riskiness for those who left their homes showed quite different accuracies between countries. The variables that appeared most frequently in the top 3 most important factors by country were the *anticipated number of people met while traveling* (i.e., 8 out of 16 countries), as well as *activity importance* (i.e., 6 countries) and *trust in people met* (i.e., 6 countries). The increases in the *anticipated number of people met while traveling* were associated with increased activity risk. The *anticipated number of people met while traveling* was the most important factor in Austria, Hungary, and the USA. This factor was particularly not important in Poland and Russia, however. *Activity importance* was a strong predictor of activity risk in most countries. Seemingly, as activity importance increased, the riskiness of the activity decreased. *Trust in people met* also had a negative relationship with the riskiness of the activity. These latter two findings suggest that individuals who leave their homes for non-essential but important activities with people they trust may be minimizing their risk-taking. Based on the root mean squared error values and R^2 values in Table 2, our models were not always accurate in predicting the risk (i.e., risk of infection) of out-of-home

activities. In countries with large sample sizes, the models were generally more accurate and accounted for more of the overall variance than in countries with relatively small sample sizes. Compared to predicting when people left their homes, however, the importance of the measured factors in predicting activity riskiness varied more widely.

While the present study explored 14 potential predictors of non-adherence to lockdown recommendations, our research is limited by the exclusion of unidentified contributors. Further, additional context-specific factors that contribute to the riskiness of an out-of-home activity may have yielded stronger or more consistent predictions than the factors we included. Our operationalization of activity risk likewise limits our conclusions. The context and sample differences between countries are also worth noting. The sample sizes, data collection methods, rates of infection, and lockdown recommendations varied between (and sometimes within) countries. The inaccurate risk score predictions might be a consequence of relatively low sample sizes in some of the countries. Also, not all countries were in a lockdown during data collection, which means that in some cases we had to rely on how the participants remembered their situation.

Overall, we can conclude that feelings of *responsibility* about the transmission of a disease is the most important predictor of adherence to lockdown recommendations, along with ,*the fear of being infected* and *feeling caged at home*. These results have important public health implications. Messaging to convince people to stay home during lockdown should appeal to personal responsibility. Perhaps, compliance could be increased with an intervention stressing that every person has an active role in a pandemic situation and that every bit of effort, even just staying at home, is a valuable and important contribution. Attempts to decrease social isolation and reframe confinement in a positive

light (e.g., as a chance for introspection) may also prove effective. A transparent and thorough coverage of symptoms, infection rates, and the possible risks that arise when contracting the disease may also help people reevaluate their priorities and motivate them to comply with confinement regulations.

Data Availability: The data and analysis script are available at https://osf.io/dfsvb/.

Funding: Balazs Aczel, Nandor Hajdu and Barnabas Szaszi were supported by the Hungarian National Research, Development and Innovation Office (NKFIH-1157-8/2019-DT); Gabriel Baník was supported by APVV-17-0418; Patrícia Arriaga was supported by the Portuguese National Funding Agency for Science and Technology (FCT, REF UID/PSI/03125/2020).; Ivan Ropovik was supported by PRIMUS/20/HUM/009; Matus Adamkovic was supported by the Slovak Research and Development Agency [project no. APVV-20-0319]; Dmitry Grigoryev and Dmitrii Dubrov were supported by the HSE University Basic Research Program; Krystian Barzykowski was supported by the National Science Centre, Poland (UMO-2019/35/B/HS6/00528).
The research reported in this paper is part of project no. BME-NVA-02, implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021 funding scheme.

Competing interests: author Martin Voracek is a PLOS ONE Editorial Board Member. This does not alter the author's adherence to PLOS ONE Editorial policies and criteria.

References

1. Yang X. Does city lockdown prevent the spread of COVID-19? New evidence from the synthetic control method. glob health res policy. 2021 Dec;6(1):20.
2. Chen R-M. Whether County Lockdown Could Deter the Contagion of COVID-19 in the USA. RMHP. 2021 Jun;Volume 14:2665–73.
3. Faulkner P. Lockdown: a case study in how to lose trust and undermine compliance. Global Discourse. 2021 May 1;11(3):497–515.
4. Hajdu N, Szaszi B, Aczel B. Extending the Choice Architecture Toolbox: The Choice Context Mapping [Internet]. PsyArXiv; 2020 Sep [cited 2021 Jul 27]. Available from: https://osf.io/cbrwt
5. Wright L, Steptoe A, Fancourt D. Predictors of self-reported adherence to COVID-19 guidelines. A longitudinal observational study of 51,600 UK adults. The Lancet Regional Health - Europe. 2021 May;4:100061.
6. UK Cabinet Office. Coronavirus restrictions: what you can and cannot do. 2021 22;
7. Stickley A, Matsubayashi T, Ueda M. Loneliness and COVID-19 preventive behaviours among Japanese adults. Journal of Public Health. 2021 Apr 12;43(1):53–60.
8. Boylan J, Seli P, Scholer AA, Danckert J. Boredom in the COVID-19 pandemic: Trait boredom proneness, the desire to act, and rule-breaking. Personality and Individual Differences. 2021 Mar;171:110387.
9. Bozdağ F. The psychological effects of staying home due to the COVID-19 pandemic. The Journal of General Psychology. 2021 Jan 5;1–23.
10. Clark C, Davila A, Regis M, Kraus S. Predictors of COVID-19 voluntary compliance behaviors: An international investigation. Global Transitions. 2020;2:76–82.
11. Bargain O, Aminjonov U. Trust and compliance to public health policies in times of COVID-19. Journal of Public Economics. 2020;192:104316.
12. Harper CA, Satchell LP, Fido D, Latzman RD. Functional Fear Predicts Public Health Compliance in the COVID-19 Pandemic. Int J Ment Health Addiction [Internet]. 2020 Apr 27 [cited 2021 May 28]; Available from:

https://link.springer.com/10.1007/s11469-020-00281-5

13. Akesson J, Ashworth-Hayes S, Hahn R, Metcalfe R, Rasooly I. Fatalism, Beliefs, and Behaviors During the COVID-19 Pandemic [Internet]. Cambridge, MA: National Bureau of Economic Research; 2020 May [cited 2021 May 28] p. w27245. Report No.: w27245. Available from: http://www.nber.org/papers/w27245.pdf

14. French Bourgeois L, Harell A, Stephenson LB. To Follow or Not to Follow: Social Norms and Civic Duty during a Pandemic. Can J Pol Sci. 2020 Jun;53(2):273–8.

15. DeFranza D, Lindow M, Harrison K, Mishra A, Mishra H. Religion and reactance to COVID-19 mitigation guidelines. American Psychologist [Internet]. 2020 Aug 10 [cited 2021 Jun 4]; Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/amp0000717

16. Alessandri G, Filosa L, Tisak MS, Crocetti E, Crea G, Avanzi L. Moral Disengagement and Generalized Social Trust as Mediators and Moderators of Rule-Respecting Behaviors During the COVID-19 Outbreak. Frontiers in Psychology. 2020;11:2102.

17. Kleitman S, Fullerton DJ, Zhang LM, Blanchard MD, Lee J, Stankov L, et al. To comply or not comply? A latent profile analysis of behaviours and attitudes during the COVID-19 pandemic. Gesser-Edelsburg A, editor. PLoS ONE. 2021 Jul 29;16(7):e0255268.

18. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.Fosu GO, Edunyah G. Flattening the exponential growth curve of covid-19 in Ghana and other developing countries; divine intervention is a necessity. Divine Interv Necessity April 2020. 2020;

Supporting information

Contextual factors predicting compliance behavior during the Covid-19 pandemic: A machine learning analysis on survey data from 16 countries

Nandor Hajdu[1,2,*], Balazs Aczel[2], Gergely Acs[3], Kathleen Schmidt[4], Jan P. Röer[5], Alberto Mirisola[6], Isabella Giammusso[6], Patrícia Arriaga[7], Rafael Ribeiro[7], Dmitrii Dubrov[8], Dmitry Grigoryev[8], Nwadiogo C. Arinze[9], Martin Voracek[10], Stefan Stieger[11], Matus Adamkovic[12,13], Mahmoud Elsherif[14], Bettina M. J. Kern[15,16], Krystian Barzykowski[17], Ewa Ilczuk[17], Marcel Martončik[12], Ivan Ropovik[18,19], Susana Ruiz-Fernandez[20,21], Gabriel Baník[12], José Luis Ulloa[22], Barnabas Szaszi[2]

[1]Doctoral School of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary, [2]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary, [3]Crysys Lab, BME-HIT, Budapest, Hungary, [4]Southern Illinois University, [5]Department of Psychology and Psychotherapy, Witten/Herdecke University, [6]Department of Psychology, Educational Science and Human Movement, University of Palermo, Italy, [7]ISCTE-University Institute of Lisbon, CIS-IUL, Portugal, [8]National Research University Higher School of Economics, Russian Federation , [9]Alex Ekwueme Federal University, Ndufu-Alike, Nigeria, [10]Department of Cognition, Emotion, and Methods in Psychology; Faculty of Psychology; University of Vienna, [11]Department of Psychology and Psychodynamics; Division Psychological Methodology; Karl Landsteiner University of Health Sciences, [12]Institute of Psychology, Faculty of Arts, University of Presov, [13]Institute of social sciences, CSPS Slovak Academy of Sciences, [14]School of Psychology, University of Birmingham, [15]Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of VIenna, [16]Department of European and Comparative Literature and Language Studies, Faculty of Philological and Cultural Studies, University of Vienna, [17]Institute of Psychology, Faculty of Philosophy, Jagiellonian University, Krakow, Poland, [18]Faculty of Education, Charles University, [19]Faculty of Education, University of Presov, [20]FOM University of Applied Sciences, [21]Leibniz-Institut für Wissensmedien, [22]Programa de Investigación Asociativa (PIA) en Ciencias Cognitivas, Centro de Investigación en Ciencias Cognitivas (CICC), Facultad de Psicología, Universidad de Talca, Chile.

* Corresponding author

Email: hajdu.nandor93@gmail.com (NH)

Methods

The separation of training and test sets were done after standardizing the variables, which might have caused data leakage. Data collection was organized by multiple laboratories, who used different methods of acquiring participants, ranging from paid platforms to social media.

Deviations from Preregistration

We preregistered our study after data collection, but before any analyses were conducted. While in our preregistration we stated that we would train random forest models on Hungarian data only, we decided to use this method on data from every country for two reasons: the robustness of the method and the comparability of results. Our preregistration also contained plans for cluster analyses, but we decided against performing them because they would not contribute to the identification of contextual factors that predict leaving home or the riskiness of the visited place - which was the main goal of the article.

Analyses

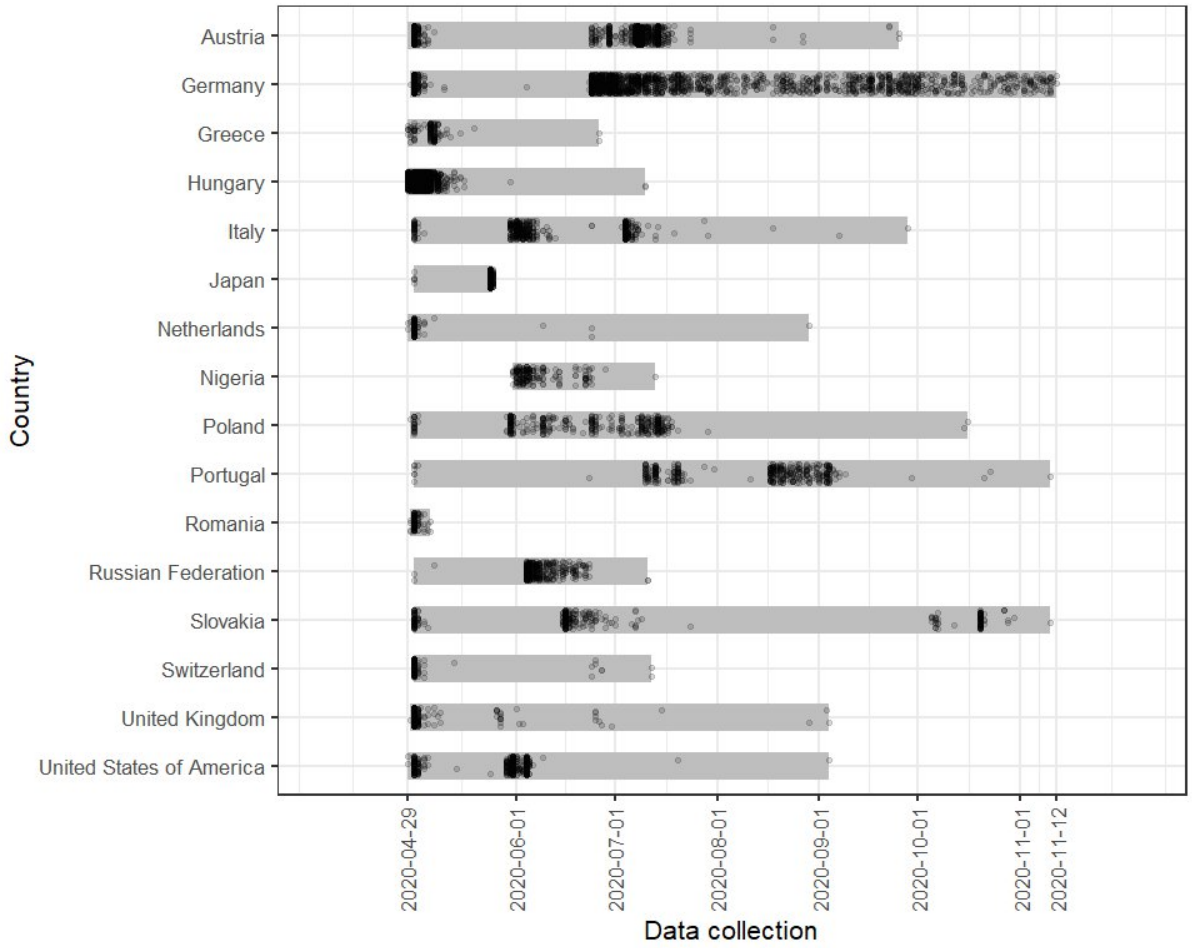
We would like to clarify that for the calculation of the random forest models, Gini impurity scores were used for splitting nodes. However, for the variance importance scores, Permutation importance scores were calculated. Only one parameter of the random forest models was tuned via cross validation: this parameter is called ‘mtry’ in the *ranger* package in R, which was used for our calculations. This parameter is the number of variables that are available to be considered at each split. We calculated different accuracy metrics for the classification of whether people reported staying at home or leaving.

Supplementary Results

S1 Table 1. Data Collection Intervals Per Countries

Country	Start date	Last participant
Austria	2020.04.29	2020.09.25
Germany	2020.04.30	2020.11.12
Greece	2020.04.29	2020.06.26
Hungary	2020.04.29	2020.07.10
Italy	2020.05.01	2020.09.28
Japan	2020.05.01	2020.05.25
Netherlands	2020.04.29	2020.08.29
Nigeria	2020.05.31	2020.07.13
Poland	2020.04.30	2020.10.19
Portugal	2020.05.01	2020.11.10
Romania	2020.04.30	2020.05.30
Russian Federation	2020.05.01	2020.07.11
Slovakia	2020.05.01	2020.11.10
Switzerland	2020.05.01	2020.07.12
United Kingdom	2020.04.30	2020.09.04
United States of America	2020.04.29	2020.09.04

S2 Figure 1. Data Collection Dates.

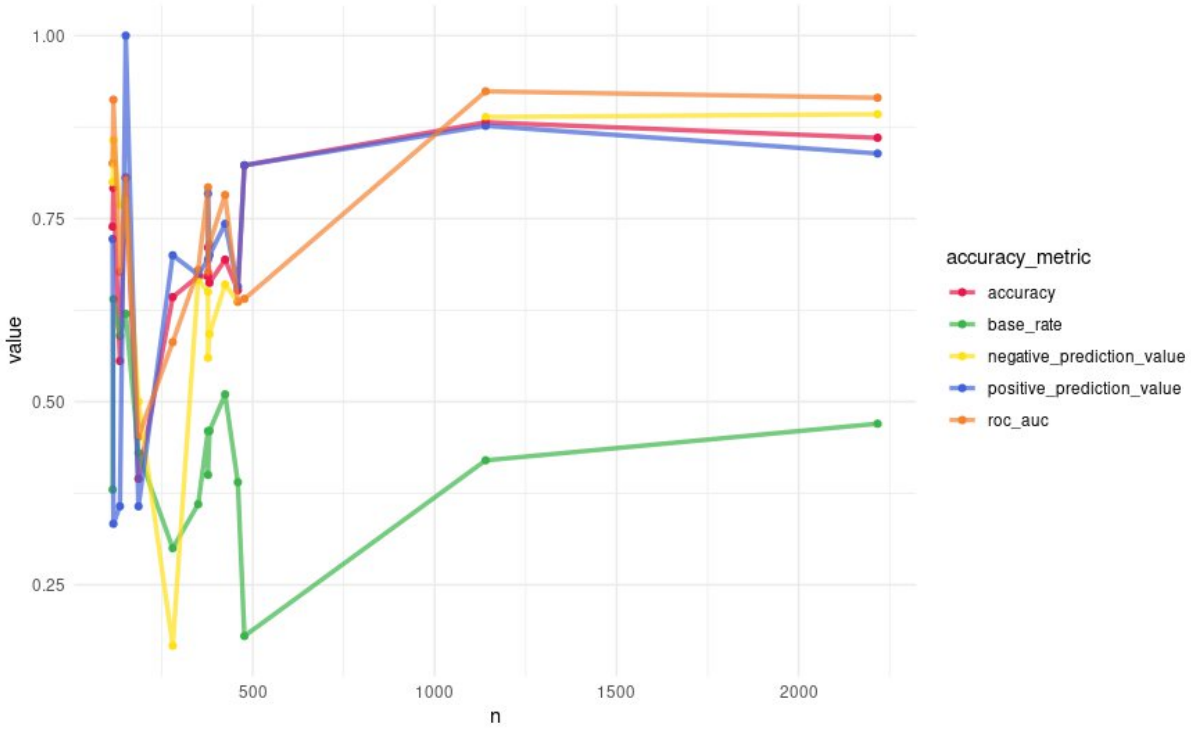


S3 Table 2. Accuracy metrics of leaving home classifiers

Country	Base rate	Accuracy	Negative prediction Value	Positive prediction Value	ROC AUC
Austria	0.42	0.8816	0.8889	0.8768	0.9239
Germany	0.47	0.8604	0.8927	0.8390	0.9151
Greece	0.59	0.5556	0.7692	0.3571	0.6776
Hungary	0.52	0.7138	0.7222	0.7050	0.7804

Italy	0.18	0.8229	-	0.8229	0.6404
Japan	0.30	0.6429	0.1667	0.7000	0.5813
Netherlands	0.64	0.7917	0.8571	0.3333	0.9125
Nigeria	0.43	0.3947	0.5000	0.3571	0.4522
Poland	0.46	0.6711	0.6500	0.6944	0.7928
Portugal	0.46	0.6623	0.5926	0.7000	0.7153
Romania	0.38	0.7391	0.8000	0.7222	0.8254
Russian Federation	0.40	0.7105	0.5600	0.7843	0.6761
Slovakia	0.36	0.6714	0.6667	0.6735	0.6800
Switzerland	0.62	0.8065	0.7778	1.0000	0.8048
United Kingdom	0.39	0.6522	0.6364	0.6571	0.6365
USA	0.51	0.6941	0.6600	0.7429	0.7824

S4 Figure 2. Accuracy metrics by sample sizes from each country.



Discussion

Overfitting can be an issue with random forest models. If the accuracy of models in many countries is lower than the base-rate, it indicates a high level of overfitting by the random forest method. In our analyses, we found that accuracy was below the base rate in only two cases: Greece and Nigeria. In these two cases, the models might have been overfitted. The Italian model predicted that everyone stayed home, which led to high accuracy, but given that only 18% of Italians left their homes, this accuracy is not an informative metric. This can be said about other accuracy measures with underlying class imbalances, in general. Our study was a retrospective study, in which the act or choice of the individual could easily influence what types of responses they give to a post-hoc survey. People often rationalize their behavior after the fact, and provide inaccurate reasons, especially when social desirability and expectations are at play.

Chapter III.

Szaszi, B., Hajdu, N., Szecsi, P., Tipton, E., & Aczel, B. (2022). A machine learning analysis of the relationship of demographics and social gathering attendance from 41 countries during pandemic. *Scientific reports*, *12*(1), 724.

Barnabas Szaszi[1*], Nandor Hajdu[1,2], Peter Szecsi[1,2], Elizabeth Tipton[3], Balazs Aczel[1]

[1] ELTE, Eotvos Lorand University, Hungary

[2] Doctoral School of Psychology, Institute of Psychology, Eotvos Lorand University, Hungary,

[3]Northwestern University, US

Correspondence: szaszi.barnabas@ppk.elte.hu

This work was completed as part of the ELTE Thematic Excellence Programme 2020 supported by the National Research, Development and Innovation Office (TKP2020-IKA-05)

Abstract

Knowing who to target with certain messages is the prerequisite of efficient public health campaigns during pandemics. Using the COVID-19 pandemic situation, we explored which facets of the society - defined by age, gender, income, and education levels - are the most likely to visit social gatherings and aggravate the spread of a disease. Analyzing the reported behavior of 87,169 individuals from and 41 countries, we found that in the majority of the countries, the proportion of social gathering-goers was higher in male than female, younger than older, lower-educated than higher educated, and low-income than high-income subgroups of the populations. However, the data showed noteworthy heterogeneity between the countries warranting against generalizing from one country to another. The analysis also revealed that relative to other demographic factors, income was the strongest predictor of avoidance of social gatherings followed by age, education and gender. Although the observed strength of these associations were relatively small, we argue that incorporating demographic-based segmentation into public health campaigns can increase the efficiency of campaigns with an important caveat: the exploration of these associations need to be done on a country level before using the information to target populations in behavior change interventions.

Keywords: social distancing, behavioral interventions, COVID-19, prevention behavior, pandemic

Introduction

When there is no medical treatment available, the best way to defy an unfolding epidemic is to convince people to adopt behavior patterns that can alleviate the spread of the disease [1,2]. Social distancing and more specifically, the avoidance of social gatherings, has been pointed out as an effective tool to decrease the spread of viruses [3–8], just as it is recommended or mandated in many countries around the world during epidemics. However, the effectiveness of these mandates greatly depends on the speed of its adoption and the level of its adherence on a societal level. In this study, we used the COVID-19 pandemic situation to explore the demographic groups that are the most likely to visit social gatherings during epidemic emergencies in order to support public-health officials and policymakers to design targeted and more efficient campaigns during epidemic emergencies.

Knowing who to target is the prerequisite of quick and efficient public health campaigns. Having information on the key populations enables policymakers to design interventions that can take into account the specific context and the characteristics of the target group. Compared to group-tailored messaging, 'one-fit-all' interventions ignore the diversity of the populations, therefore, they are expected to be less efficient, potentially on the cost of human lives [9]. Taking a step further, it is possible that during pandemics, the same behavioral intervention has opposite effects on different populations. For example, the same campaign can increase adherence behavior in one group and motivate non-adherence in another [e.g., 10] leading to avoidable death. As previous research suggests that the majority of people comply with the social distancing recommendations [11], to avoid unintended detrimental consequences, it is crucial that policymakers only target public health campaigns on those groups whose behavior needs to be changed. Just as doctors do not perform operations on healthy people.

While intervention designers can target their audience based on age, gender, income, and education through the majority of the communication platforms, no information is available about the risk perception or the norms of individuals neither on television, print nor online channels. That is, although latent factors such as values, norms, risk-perception may have great explanatory power on social distancing behavior [e.g., 12–14], in the present research we focused on the demographic factors which are widely available to use for targeting public health campaigns and policies.

Accordingly, we identify the social gathering goers based on age, education, gender, and income. Previous research showed that adherence behavior during epidemics significantly varies along with these demographic factors [for a review see 15]. However, prior studies were mainly conducted in one country at a time and the association of these variables with social distancing also showed a mixed picture. Age has been positively associated with preventive behaviors during pandemics in some studies [e.g., 16–19], and negatively in others [20], while some studies found no association [e.g., 21,22]. Income has been mostly found negatively correlated with non-complying behavior [21,23–26], a result which is argued to be found because low-paid workers are less able to work remotely and stay at home without losing their jobs. Studies investigating the relationship of gender and protective behavior also found mixed results: although most studies found that men are less likely than women to adhere to the protective recommendations [19,e.g., 27–29], other studies showed evidence for the opposite [16] or found no association [18]. Educational level has also been associated with opposing results: while some studies showed that higher education predicts more precautionary behavior [30,31], others found contrary evidence [22] or mixed results [19].The diversity of these results might be due to the typically small sample sizes or might reflect the fact that the samples come from one or a few number of varying countries with diverse populations. The small samples and their geographical dispersion also make it hard to generalize from these former findings.

In the present research, extending previous results focusing on a handful of countries, we explore the association between the demographic factors and the avoidance of social gatherings in a large sample (more than 80,000 individuals from 41 countries) which makes the results generalizable and comparable across different cultures. We use a machine-learning technique which allows us to identify not just the main effects of the demographic factors but also reveal the subtle patterns and explore the heterogeneity between the countries. We also discuss how these results can be used to improve public health interventions.

Methods

Dataset

The dataset was collected by an international research group during the early phase of the COVID-19 pandemic between 2020.03.20. and 2020.04.07. [32]. The data were gathered via snowball method using an online survey; participants were recruited

from all over the world via several social media and media outlets. As a result, 112,136 individuals filled out the survey from 175 countries. For the detailed method of data collection, the full list of items and their order consult 32.

From this dataset, we excluded the responses of individuals who did not complete the full survey. Furthermore, we included responses only from those countries, where, at the time of survey compilation, the country of residence of the participant has adopted some form of restriction or recommendation affecting social gatherings. That is, we checked whether the Government of each country had any active measure on social gatherings, public events, workplaces, public transport, schools or internal movement in the respondents country using the data from the Oxford COVID-19 Government Response Tracker [33], which collects publicly available information on COVID-19 related governmental responses in each country. Responses with nonsensical values were removed (age < 99; household members = 0; 4 < years of education < age - 5) as well as individuals reporting income greater than 99 percent of the sample within their country. Finally, to maximize the reliability of the survey in each country, responses from countries with fewer than 400 respondents were also not analyzed. As a result, our final dataset consisted of 87,169 individual responses from 41 countries (56% female, M_{age} = 40.0, SD_{age} = 12.8) with 2126 mean number of respondents per country. As of 2020, these countries accounted for 73.05% of the world's population[1]. A detailed description of the sample in each country can be found in Supplementary Table 1. The data are available at the projects' OSF page: https://osf.io/rehc7/.

Procedures and Measures

As part of a broader online data collection effort (Fetzer et al. 2020), participants responded to several COVID-19 related survey items. Crucially, for the purposes of this study, respondents were asked to indicate on a 100 point scale to what extent the statement '*I did not attend social gatherings*' describes their behavior for the past week. This item was our key measure assessing individuals' behavior regarding social gatherings. We categorized individuals who indicated total agreement (*100 points*) with the statement that they did not attend social gatherings as *social gathering avoiders*, while the rest of the participants were classified as *social gathering goers*.

[1]World population estimation was based on the UN's World Population Prospects, accessed from Word Population Review [34]

Furthermore, participants responded to several questions regarding their demographics including age (*Which year were you born?),* gender (*Which gender do you identify with? Male; Female; Other*), education (*How many years of education did you complete?*), country of residence (*In which country do you mostly live?*). Participants also indicated their household income (*What is your monthly household income, before tax, your country's*) and their household size *(How many people live in your household?*). Following previous recommendations [35], we used adjusted household income in our analyses. Adjusted household income was calculated by dividing household income by the square root of household size.

Data analysis strategy

To explore the role of demographics in the avoidance of social gatherings, random forest models were applied as they are robust to the non-linearity of the data and handle unbalanced data relatively well compared to logistic regression models[2].

Instead of fitting a global model on the overall population, we fit individual models to each country, as the disease progression, policy measures, and political and public health messaging -- as well as more general social and behavioral norms -- vary dramatically from country to country in ways that are difficult to appropriately adjust and control for. Moreover, using county specific models enables us to explore the heterogeneity among the countries. Note, that although we access a relatively large sample from each country, we re-weight the observations based on the respondent's gender, age, income, and education in the main analyses to make the collected data more representative at the country level.

In our analyses, we split data from each country into training and test sets in an 80-20 ratio, and we use the training set to find the number of variables sampled at each split of a decision tree for our random forest models. The number of variables is set between 1 and 4 and tuned separately for every country via repeated 10-fold cross-validation. Because the number of social gathering goers (24%) and avoiders (76%) is unequal in our dataset, we upsample the training data; this means that we sample with replacement from the original, minority class data until we reach a sample size equal to the majority class. This way, in the training data, social gathering avoiders and goers are

[2] Random forest models operate by creating decision trees. First, the independent variable that best separates the categories of the dependent variable is selected, and data is separated into two groups based on this variable. Next, for each group, the method picks the independent variable that best separates the data. This process is repeated until the predefined tree-depth is reached [36].

balanced. From the many well-established accuracy metrics, we chose to tune our models on the training set to get the greatest area under the precision-recall curve (prAUC), because it is fairly robust to unbalanced data. Finally, we use the test set to see how well each previously trained model performs. The analysis code is available at https://osf.io/rehc7/.

Results

We created and ran random forest models for each country with the specifications detailed above. The models successfully predicted attendance of social gatherings based on demographic factors during the earthly phase of the pandemics but also showed significant cross-country heterogeneity ranging from 0.52 to 0.84. For a detailed description of the models and prediction accuracies see Supplementary Table 2.

The association between the demographic factors and the avoidance of social gatherings across countries

First, to explore the association of demographic variables and the avoidance of social gathering across the world, we calculated descriptive statistics on the proportion of social gathering goers and avoiders in each country in the following subgroups: female participants, male participants; individuals reporting lower than the median income, higher than the median income; lower than median age, higher than median age; lower than median education, and higher than median education. Based on this categorization, we found that the proportion of individuals violating social distancing was higher for males than females in 95% of the countries (39 countries), among low-income than high-income people in 80% of the countries (33 countries), among younger than older people in 78% of the countries (32 countries), and among lower educated than higher educated people in 66% of the countries (27 countries).

Next, using the results from the random forest models, we created partial dependence plots in order to see whether each of the demographic factors were associated with higher or lower probability of social gathering avoidance (Figure 1). Partial dependence plots show the average predicted probability of leaving home associated with a given value of the demographic factor in each country. Plotting these lines on the same graph for each country makes it possible to recognize mutual trends in the change of

probabilities and explore the heterogeneity of the results. Accordingly, Figure 1 shows that in most of the countries, being at older age, being female, having a lower income, and more years of education seem to indicate a lower probability of attending social gatherings, although there is significant heterogeneity across countries for each of the demographics variables except gender.

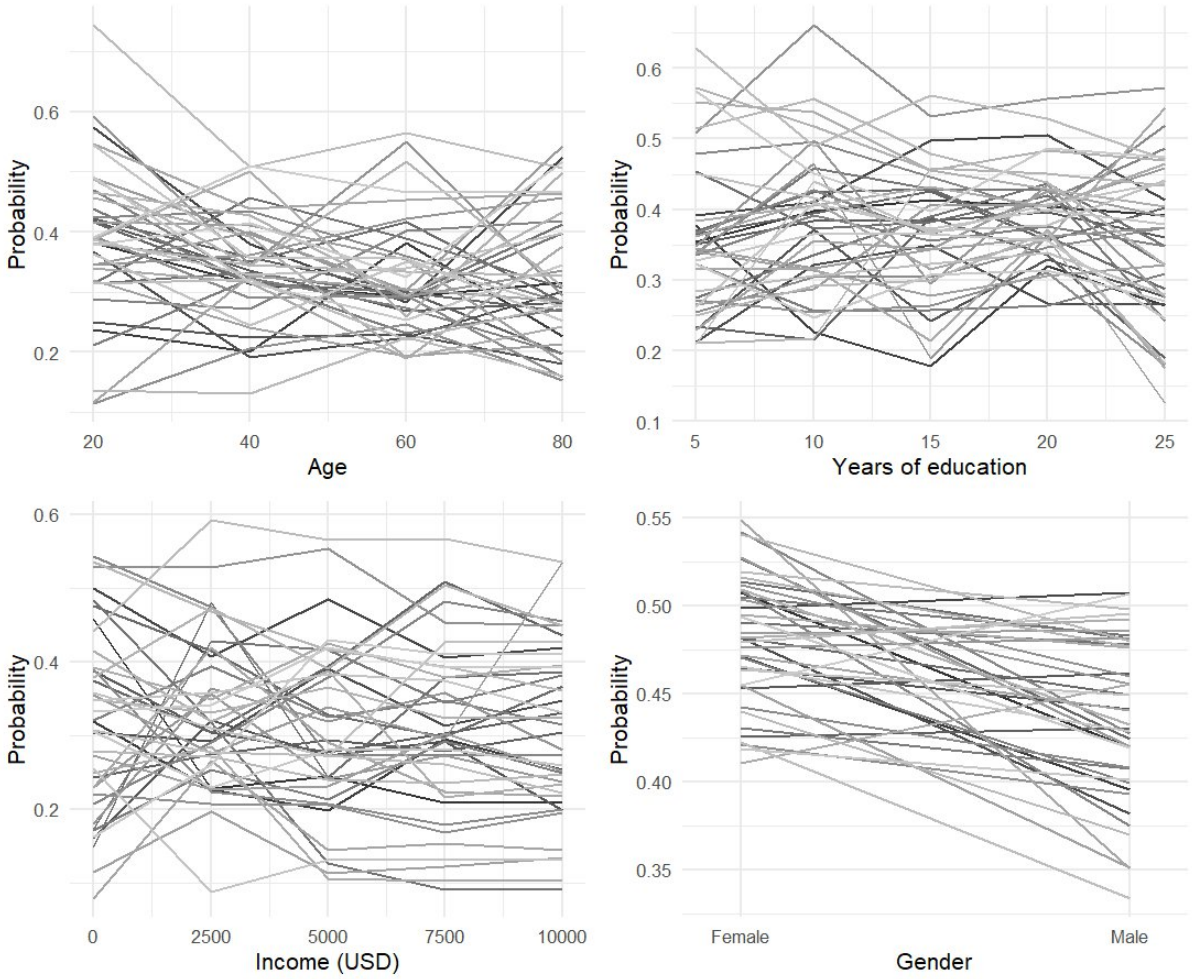


Figure 1 Partial dependence plots show the average predicted probability of leaving home associated with a given value of the demographic factor of age, years of education, income, and gender (in different plots) for all the countries.

General importance of demographic factors at predicting the avoidance of social gatherings

We also calculated variable importance scores for each demographic factor in each country. The variable importance score is a metric expressing the mean increase in

accuracy when a given variable is added to a model, that is, it shows how important a variable is at improving the overall predictive power of a model with the other parameters keeping constant.

The median importance score of income was 0.07 (with a range of 0 - 0.23), meaning that adding information about income would make our predictions around 7% more accurate. This value was 0.05 (with a range of 0 - 0.21) in the case of age, 0.05 (with a range of 0 - 0.23) in the case of education, and 0.02 (with a range of 0 - 0.13) in the case of gender. Figure 2 summarizes the variable importance scores for each demographic factor in each country.

Relative importance of demographic factors at predicting the avoidance of social gatherings

We determined the strongest predictor of social gathering avoidance by determining the demographic factor with the greatest variable importance score in each country. Out of the 41 countries examined, the strongest predictor was income in 29 countries, age in 10 countries, and education in 2 countries. The variable with the lowest importance score was gender in 36 countries, years of education in 4 countries, and income in 1 country (see Figure 1).

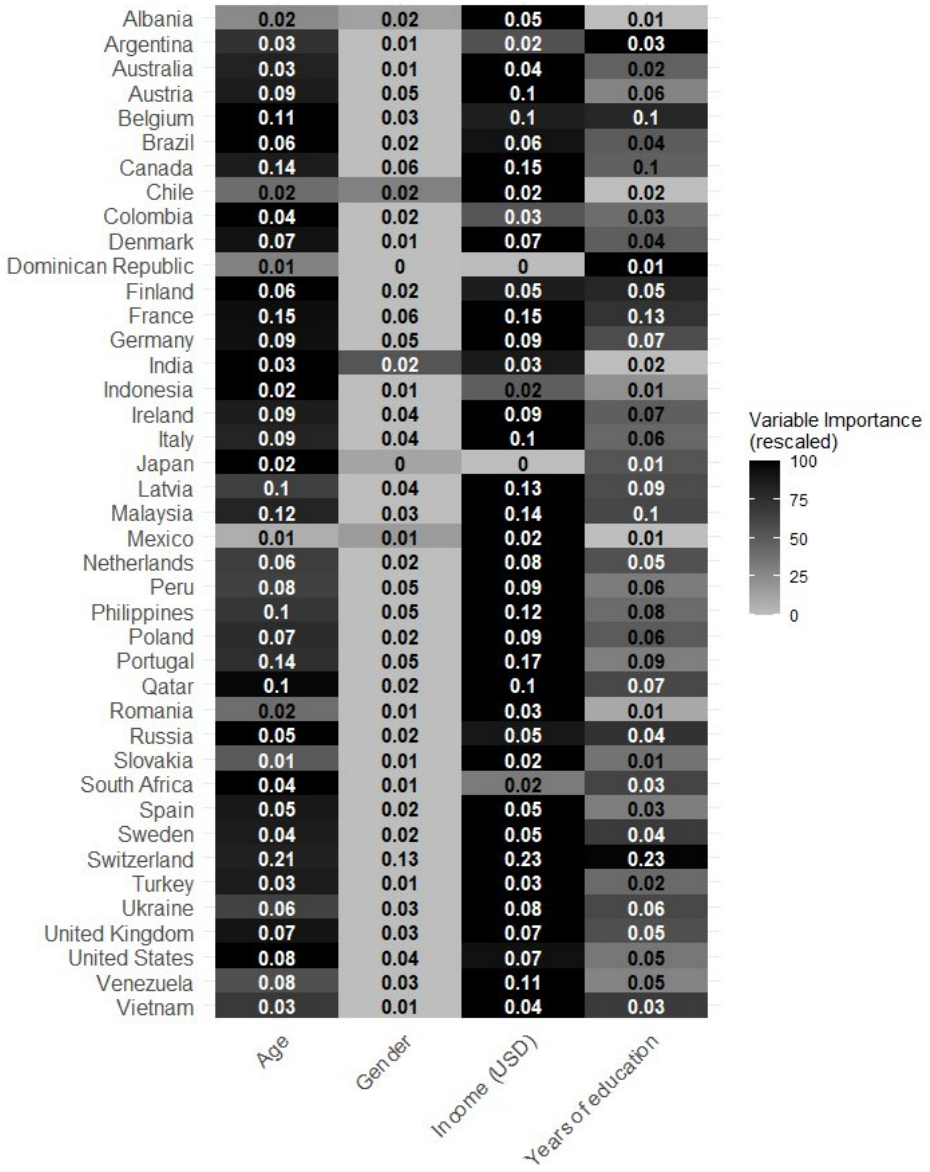


Figure 2. The figure summarizes the variable importance scores for each demographic variable in each country. Variable importance values express the mean increase in accuracy when a given demographic variable is added to a model. The coloring of the figures depicts the relative importance of the variables within each country while the variable importance values were rescaled between 0 and 100 in each country, 100 being the most (darkest) and 0 being the least important (lightest).

Discussion

To explore which demographic subgroups are the most likely to visit social gatherings during epidemic emergencies, we investigated a large dataset collected during the early phase of the COVID-19 pandemic situation from 41 countries. With these

countries accounting for 73.05% of the world's population, our study provides the first global, systematic investigation of demographic factors on social distancing.

The results show that in the majority of the countries, the proportion of social-goers was higher in male than female, younger than older, lower-educated vs. higher educated, and low-income vs. high-income subgroups of the population. However, we also observed noteworthy heterogeneity among the countries regarding the direction of the association of the investigated demographic factors and the propensity to visit social gatherings. For example, in 33% of the countries, high-educated citizens were more prone to non-adhere to social distancing than their low-educated counterparts. Such heterogeneity warrants policymakers and researchers to simply generalize social distancing behavior from one country to another without understanding the specific context, and suggests that simply targeting older, low-income, low-educated, male citizens in public health campaigns is not a proper solution.

When the resources are tight and the targeting of an intervention needs to be made based on one given demographic variable, one needs to know which variable this should be. Relative to the other demographic factors, our results provided evidence that income was the strongest predictor worldwide when it comes to visiting social gatherings followed by age, education, and finally gender, but again we found large heterogeneity between the countries. Even countries that are geographically and culturally close (such as Germany and Austria) showed different patterns**.** One potential reason behind this variability is that the identification of the strongest demographic predictor can be sensitive to the correlation among the demographic variables and it also fails to account for synergistic effects between two or more demographic factors. That is, instead of finding an emerging trend across countries, the results confirmed that the investigated associations are heterogeneous, largely differing from country to country. Such findings bring further evidence that context has a non-ignorable moderating power on the relationship of social distancing and demographic factors, and they suggest that the exploration of this association needs to be done on a country level.

Although the observed strength of the associations between the demographics and avoidance of social gatherings are often small, these small effects can have meaningful and important consequences. When analyzing the country-level data, we found for example that across the investigated countries, the youngest 20% of the

population were on average ~4% more likely not to adhere to social distancing than the oldest 20% of the population (see Supplementary Materials). Note, that this is a sizable difference. Previous evidence suggests that every 1 % increase in non-essential visits lead to 7-8 % increase in new COVID-19 cases the following week [37].

Targeting these less avoidant demographic subgroups of the population with public health campaigns could have important advantages. First, affecting only those subgroups who need to be affected could save public resources and decrease the risk of the potential conflicting effects on adherent subgroups. Note, that we also found that 75% of our sample reported absolute avoidance of social gatherings, and these populations don't need to be addressed by prevention campaigns. Second, the identification of non-adherent groups would enable intervention designers to increase the effectiveness of public health campaigns by tailoring the messages to the target population-specific habits, beliefs, and attitudes.

The present study has several limitations. First, our data that we analysed were collected during the early phase of the pandemic, and it is possible that the adherence of different demographic groups changed in its later stages. Although our data are not suitable to resolve this concern, a longitudinal study found that the strength of association between different demographic groups and social distancing was similar from April to August 2020 [19]. Second, the social distancing data used in the present research are based on self-reports. The results from Gollwitzer et al. [38] suggest that this is not necessarily a problem: the authors connected self-reports with 17 million smartphone GPS coordinates during the COVID-19 pandemic and found that self-reported data followed actual social distancing behavior. Third, it needs to be noted that pandemics may vary in features driving social distancing decisions (e.g., different death rates between different demographic groups), therefore, our findings may not generalize to all future pandemics. Fourth, our study focused on the association of demographic variables and social distancing behavior but did not provide causal explanations behind the observed patterns. Future studies addressing why the revealed associations emerge (e.g., differences in home working opportunities or housing conditions, perceptions, beliefs), might be able to explain some of the variance observed across the countries.

References

1. Betsch, C. *et al.* Social and behavioral consequences of mask policies during the COVID-19 pandemic. *Proc. Natl. Acad. Sci.* **117**, 21851–21853 (2020).
2. Van Bavel, J. J. *et al.* Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).
3. Cheetham, N. *et al.* Determining the level of social distancing necessary to avoid future COVID-19 epidemic waves: a modelling study for North East London. *Sci. Rep.* **11**, 1–10 (2021).
4. Chu, D. K. *et al.* Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet* (2020).
5. Cot, C., Cacciapaglia, G. & Sannino, F. Mining Google and Apple mobility data: Temporal anatomy for COVID-19 social distancing. *Sci. Rep.* **11**, 1–8 (2021).
6. Fong, M. W. *et al.* Nonpharmaceutical Measures for Pandemic Influenza in Nonhealthcare Settings-Social Distancing Measures. *Emerg. Infect. Dis.* **26**, (2020).
7. Glass, R. J., Glass, L. M., Beyeler, W. E. & Min, H. J. Targeted social distancing designs for pandemic influenza. *Emerg. Infect. Dis.* **12**, 1671 (2006).
8. Rashid, H. *et al.* Evidence compendium and advice on social distancing and other related measures for response to an influenza pandemic. *Paediatr. Respir. Rev.* **16**, 119–126 (2015).
9. Tipton, E., Bryan, C. J. & Yeager, D. S. To change the world, behavioral intervention research will need to get serious about heterogeneity. *Manuscr. Prep. Retrieved Httpsstatmodeling Stat Columbia Eduwp-Contentuploads202007Heterogeneity-1-23-20-NHB Pdf* (2020).
10. Lilienfeld, S. O. Psychological treatments that cause harm. *Perspect. Psychol. Sci.* **2**, 53–70 (2007).
11. Moore, R. C., Lee, A., Hancock, J. T., Halley, M. & Linos, E. Experience with social distancing early in the COVID-19 pandemic in the United States: Implications for Public Health Messaging. *medRxiv* (2020).
12. Clark, C., Davila, A., Regis, M. & Kraus, S. Predictors of COVID-19 voluntary compliance behaviors: An international investigation. *Glob. Transit.* **2**, 76–82 (2020).
13. Danckert, J., Boylan, J., Seli, P. & Scholer, A. Boredom and rule breaking

during COVID-19. (2020).
14. Hajdu, N., Aczel, B. & Szaszi, B. Factors behind home-confinement during pandemics: a machine learning approach. *Manuscript in prepration* (2021).
15. Bish, A. & Michie, S. Demographic and attitudinal determinants of protective behaviours during a pandemic: A review. *Br. J. Health Psychol.* **15**, 797–824 (2010).
16. Jones, J. H. & Salathe, M. Early assessment of anxiety and behavioral response to novel swine-origin influenza A (H1N1). *PLoS One* **4**, e8032 (2009).
17. Leung, G. M. *et al.* Longitudinal assessment of community psychobehavioral responses during and after the 2003 outbreak of severe acute respiratory syndrome in Hong Kong. *Clin. Infect. Dis.* **40**, 1713–1720 (2005).
18. Megreya, A. M., Latzman, R. D., Al-Ahmadi, A. M. & Al-Dosari, N. F. The COVID-19-Related Lockdown in Qatar: Associations Among Demographics, Social Distancing, Mood Changes, and Quality of Life. *Int. J. Ment. Health Addict.* 1–17 (2021).
19. Reinders Folmer, C. *et al.* Compliance in the 1.5 Meter Society: Longitudinal Analysis of Citizens' Adherence to COVID-19 Mitigation Measures in a Representative Sample in the Netherlands in Early April, Early May, and Late May. *Early May Late May June 11 2020* (2020).
20. Brug, J. *et al.* SARS risk perception, knowledge, precautions, and information sources, the Netherlands. *Emerg. Infect. Dis.* **10**, 1486 (2004).
21. Papageorge, N. W. *et al. Socio-Demographic Factors Associated with Self-Protecting Behavior during the COVID-19 Pandemic*. (2020).
22. Quinn, S. C., Kumar, S., Freimuth, V. S., Kidwell, K. & Musa, D. Quinn, S. C., Kumar, S., Freimuth, V. S., Kidwell, K., & Musa, D. (2009). Public willingness to take a vaccine or drug under Emergency Use Authorization during the 2009 H1N1 pandemic. *Biosecurity and bioterrorism: biodefense strategy, practice, and science,* **7**, 275–290 (2009).
23. Baum, N. M., Jacobson, P. D. & Goold, S. D. "Listen to the people": public deliberation about social distancing measures in a pandemic. *Am. J. Bioeth.* **9**, 4–14 (2009).
24. Blake, K. D., Blendon, R. J. & Viswanath, K. Employment and compliance with pandemic influenza mitigation recommendations. *Emerg. Infect. Dis.* **16**, 212 (2010).

25. Garnier, R., Benetka, J. R., Kraemer, J. & Bansal, S. Socioeconomic disparities in social distancing during the COVID-19 pandemic in the United States: observational study. *J. Med. Internet Res.* **23**, e24591 (2021).
26. Weill, J. A., Stigler, M., Deschenes, O. & Springborn, M. R. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc. Natl. Acad. Sci.* **117**, 19658–19660 (2020).
27. Barr, M. *et al.* Pandemic influenza in Australia: using telephone surveys to measure perceptions of threat and willingness to comply. *BMC Infect. Dis.* **8**, 117 (2008).
28. Lau, J. T. F., Yang, X., Tsui, H. & Kim, J. H. Monitoring community responses to the SARS epidemic in Hong Kong: from day 10 to day 62. *J. Epidemiol. Community Health* **57**, 864–870 (2003).
29. Pedersen, M. J. & Favero, N. Social Distancing During the COVID-19 Pandemic: Who Are the Present and Future Non-compliers? *Public Adm. Rev.* (2020).
30. Leung, G. M. *et al.* The impact of community psychological responses on outbreak control for severe acute respiratory syndrome in Hong Kong. *J. Epidemiol. Community Health* **57**, 857–863 (2003).
31. Leung, G. M. *et al.* A tale of two cities: community psychobehavioral surveillance and related impact on outbreak control in Hong Kong and Singapore during the severe acute respiratory syndrome epidemic. *Infect. Control Hosp. Epidemiol.* **25**, 1033–1041 (2004).
32. Fetzer, T. *et al.* Global behaviors and perceptions in the COVID-19 pandemic. (2020).
33. Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5**, 529–538 (2021).
34. World Population Review. World Population Review. Retrieved from https://www.worldometers.info/world-population/population-by-country/ on 09.11.2020. (2020).
35. Organisation for Economic Co-operation and Development,. *Divided we stand: Why inequality keeps rising*. (OECD Paris, 2011).
36. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
37. Sharkey, P. & Wood, G. The Causal Effect of Social Distancing on the Spread of SARS-CoV-2. (2020).

38. Gollwitzer, A., Martel, C., Marshall, J., Höhs, J. M. & Bargh, J. A. Connecting self-reported social distancing to real-world behavior at the individual and us state level. (2020).
39. Aczel, B. *et al.* A consensus-based transparency checklist. *Nat. Hum. Behav.* **4**, 4–6 (2020).
40. International Monetary Fund,. International Financial statistics. Data set retrieved from https://data.imf.org/regular.aspx?key=61545862 D. (2020).

Acknowledgements

We would like to thank Jared Murray for his insights and analyses on the initial version of this manuscript and Melinda Szrenka for her supporting love and patience throughout the process.

Author contributions

Conceptualization: B.S, B.A, N.H., and P. S.; Methodology: B.S, N.H., and P. S.;Project Administration: B.S.; Supervision: B.A, E.T.; Writing - Original Draft Preparation: B.A; Writing - Review & Editing: B.S, B.A, N.H., E.T., and P. S.;

Competing Interests

The author(s) declare no competing interests.

Openness Statement

All the data and analysis code of this project are available from https://osf.io/rehc7/**.** The transparency report[39] of the project is available from https://osf.io/f3nug/

Figure legends

Figure 1 Partial dependence plots show the average predicted probability of

leaving home associated with a given value of the demographic factor of age, years of education, income, and gender (in different plots) for all the countries. Figure 2 The figure summarizes the variable importance scores for each demographic variable in each country. Variable importance values express the mean increase in accuracy when a given demographic variable is added to a model. The coloring of the figures depicts the relative importance of the variables within each country while the variable importance values were rescaled between 0 and 100 in each country, 100 being the most (darkest) and 0 being the least important (lightest).

Tables

There are no tables in the main text.

Supplementary Materials

Analyses

The upsampling and the re-weighting procedure used in this project can create biases. There can be people from a rare demographic whose response weight is largely multiplied, while them being a rare demographic in the sample might be a non-typical member of said demographic to begin with. This is an important limitation of our presented research. Splitting the data into training and test sets was done after calculating descriptive statistics, but no transformation was done on the data. Upsampling was only used on the training set; we re-weighted the test set, but did not upsample it. Variable importance scores presented in the article are permutation importance scores that show a percentage of increase in prAUC. We re-calculated the random forest models without any weighting or upsampling, neither on the training, nor the test sets. Also, we fitted unweighted logistic regression models for the purpose of comparing them to the random forest models in terms of accuracy.

Results

Table 1. Descriptive Statistics Per Country

Country	**N**	**Ratio of Females**	**Ratio of Social gathering goers**	**Median adjusted household income (USD)**	**Median age**	**Median years of education**
Albania	671	0.64	0.21	486.19	34	17
Argentina	807	0.43	0.11	965.63	37	18

Australia	888	0.67	0.39	4735.95	42	17
Austria	1056	0.53	0.15	3098.82	38	17
Belgium	512	0.54	0.12	3795.27	37	17
Brazil	9351	0.61	0.23	867.35	34	17
Canada	2615	0.65	0.14	9985.2	43	18
Chile	503	0.55	0.21	2008.74	40	18
Colombia	1682	0.48	0.23	1107.06	36	18
Denmark	422	0.51	0.24	3612.84	37.5	17
Dominican Republic	472	0.52	0.19	1575.79	37	19
Finland	725	0.54	0.29	4260.88	40	17
France	2067	0.53	0.11	3162.72	36	7
Germany	9686	0.49	0.24	3286.8	37	17
India	837	0.39	0.23	1038.16	33	18
Indonesia	1417	0.56	0.4	176.38	27	16
Ireland	668	0.49	0.13	4648.24	41	18
Italy	1744	0.47	0.09	2739	37	18

Japan	556	0.41	0.48	5311.41	45	18
Latvia	582	0.72	0.14	1095.6	34	17
Malaysia	480	0.58	0.17	1150.44	38	17
Mexico	860	0.53	0.27	1052.59	42	19
Netherlands	1297	0.57	0.23	3834.6	40	18
Peru	1835	0.42	0.11	1066.85	38	17
Philippines	621	0.68	0.13	630.62	31	16
Poland	462	0.56	0.22	1364.22	35	17
Portugal	537	0.69	0.11	1095.6	37	17
Qatar	1016	0.7	0.27	3071.52	29	16
Romania	788	0.65	0.27	1575.2	37	17
Russia	3110	0.4	0.35	668.47	33	15
Slovakia	597	0.53	0.34	1643.4	34	17
South Africa	496	0.71	0.3	1778.54	42	16
Spain	2089	0.5	0.11	2324.12	44	20
Sweden	5461	0.69	0.62	3070.49	46	16
Switzerland	3486	0.55	0.13	6592.05	41	15

Turkey	2773	0.53	0.25	459.99	31	17
Ukraine	1367	0.72	0.21	377.98	28	15
United Kingdom	10550	0.51	0.26	6992.15	43	17
United States	10686	0.61	0.18	18500	40	18
Venezuela	626	0.45	0.16	52.21	54	18
Vietnam	771	0.77	0.31	345.5	21	13

To get a comparable estimate of household incomes across different countries, we converted the declared income data to USD. As an exchange rate, we used the national currency per U.S. end of month dollar rate of March 2020 [40].

Table 2. Parameters of the tuned, weighted models and accuracy on test data.

Country	**No. of vars sampled at each split**	**prAUC**	**Accuracy**
Albania	2	0.977	0.617
Argentina	2	0.979	0.820
Australia	2	0.979	0.520
Austria	2	0.979	0.738
Belgium	2	0.978	0.802
Brazil	2	0.977	0.640
Canada	2	0.977	0.711
Chile	2	0.977	0.700
Colombia	2	0.976	0.627
Denmark	2	0.978	0.655
Dominican Republic	2	0.978	0.713
Finland	2	0.979	0.531
France	2	0.978	0.772
Germany	2	0.980	0.605
India	2	0.975	0.663

Indonesia	2	0.977	0.572
Ireland	2	0.979	0.722
Italy	2	0.970	0.842
Japan	2	0.974	0.532
Latvia	2	0.975	0.696
Malaysia	2	0.978	0.750
Mexico	2	0.978	0.673
Netherlands	2	0.976	0.667
Peru	2	0.976	0.730
Philippines	2	0.977	0.610
Poland	2	0.976	0.761
Portugal	2	0.979	0.802
Qatar	2	0.977	0.599
Romania	2	0.975	0.637
Russia	2	0.975	0.585
Slovakia	2	0.980	0.585
South Africa	2	0.980	0.626
Spain	2	0.977	0.734
Sweden	2	0.978	0.549

Switzerland	2	0.979	0.789
Turkey	2	0.976	0.596
Ukraine	2	0.982	0.621
United Kingdom	2	0.977	0.585
United States	2	0.980	0.666
Venezuela	2	0.978	0.750
Vietnam	2	0.979	0.569

Number of variables sampled at each split of a decision tree and area under the precision-recall curve of the tuned models by country. Accuracy represents the prediction accuracy of each tuned model on test data.

Calculating the differences in avoidance of social gatherings between the top 20% and the bottom 20% of the population

We also calculated the difference in avoidance of social gatherings between the top 20% and the bottom 20% for each demographic factor across the investigated countries. We observed that the youngest 20% of the population were on average 4.17% (with median of 5%) more likely not to adhere to social distancing than the oldest 20% of the population, while males were on average 4.07% (with a median of 4%) more likely not to adhere to social distancing than females. We found that across the investigated countries, the poorest 20% of the population were on average 0.48%(with a median of 1%) more likely not to adhere to social distancing than the richest 20% of the population while the most educated 20% of the population were on average 3.46% (with a median of 3%) more likely not to adhere to social distancing than the least educated 20% of the population.

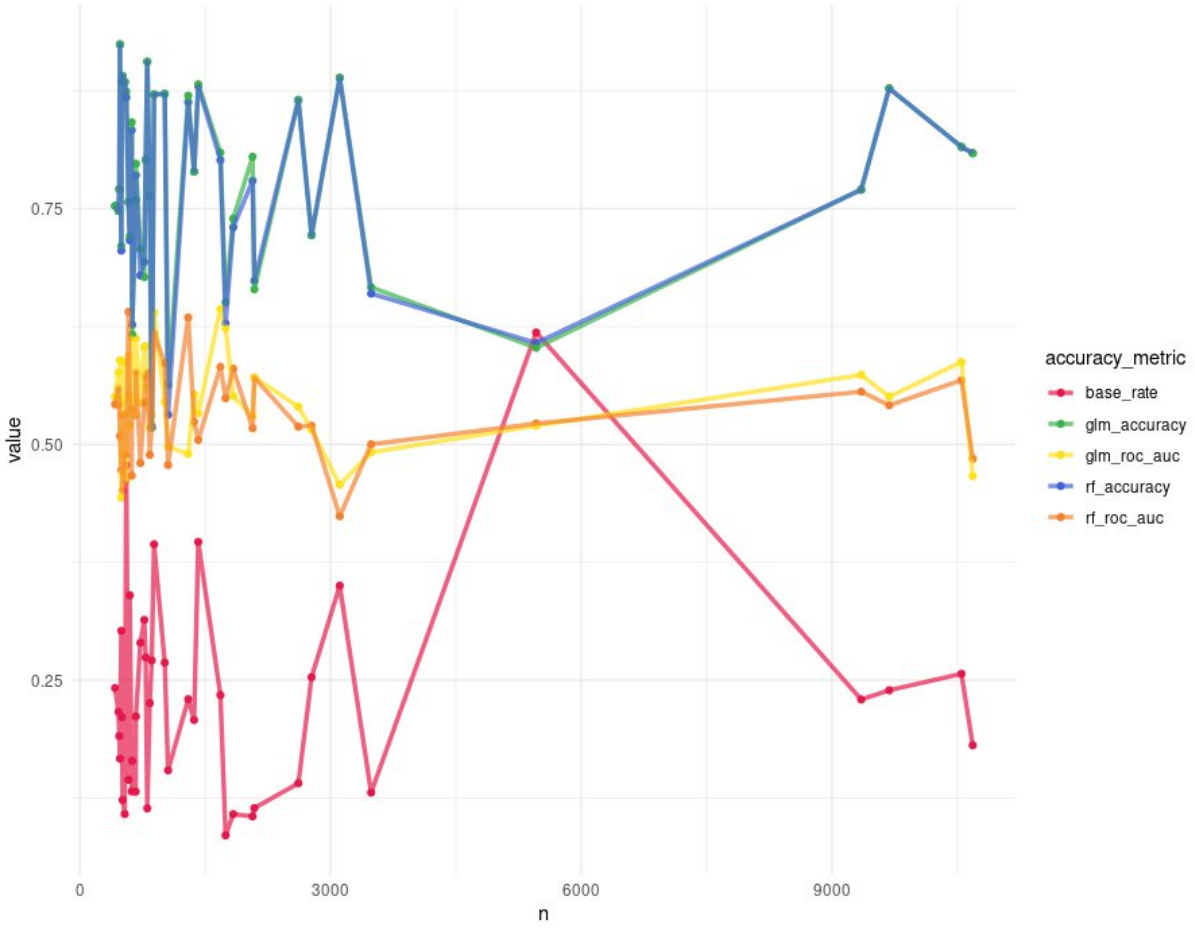
Table 3. Accuracy metrics of the unweighted, not upsampled models.

country	N	base rate	RF acc	RF npv	RF ppv	RF ROC AUC	GLM acc	GLM npv	GLM ppv	GLM ROC AUC
Albania	671	0.21	0.79	0.25	0.80	0.58	0.80	-	0.80	0.61
Argentina	807	0.11	0.91	-	0.91	0.58	0.91	-	0.91	0.56
Austria	888	0.39	0.87	-	0.87	0.62	0.87	-	0.87	0.64
Australia	1056	0.15	0.53	0.40	0.58	0.48	0.56	0.31	0.58	0.50
Belgium	512	0.12	0.89	-	0.89	0.45	0.89	-	0.89	0.49
Brazil	9351	0.23	0.77	-	0.77	0.56	0.77	-	0.77	0.57
Canada	2615	0.14	0.87	-	0.87	0.52	0.87	-	0.87	0.54
Switzerland	503	0.21	0.88	-	0.88	0.53	0.88	-	0.88	0.55
Chile	1682	0.23	0.80	0.00	0.81	0.58	0.81	-	0.81	0.64
Colombia	422	0.24	0.75	-	0.75	0.54	0.75	-	0.75	0.55
Germany	472	0.19	0.77	-	0.77	0.54	0.77	-	0.77	0.55
Denmark	725	0.29	0.68	0.00	0.70	0.48	0.71	-	0.71	0.54
Dominican Republic	2067	0.11	0.78	0.20	0.81	0.52	0.81	-	0.81	0.53
Spain	9686	0.24	0.88	-	0.88	0.54	0.88	-	0.88	0.55
Finland	837	0.23	0.76	-	0.76	0.49	0.76	-	0.76	0.52
France	1417	0.40	0.88	0.00	0.88	0.50	0.88	-	0.88	0.53
United Kingdom	668	0.13	0.76	-	0.76	0.53	0.76	-	0.76	0.53
Indonesia	1744	0.09	0.63	0.46	0.67	0.55	0.65	0.52	0.69	0.62
Ireland	556	0.48	0.87	0.33	0.88	0.49	0.87	-	0.87	0.46
India	582	0.14	0.76	-	0.76	0.64	0.76	-	0.76	0.59
Italy	480	0.17	0.92	-	0.92	0.51	0.92	-	0.92	0.59
Japan	860	0.27	0.52	0.48	0.55	0.52	0.52	0.48	0.54	0.53
Latvia	1297	0.23	0.86	0.00	0.87	0.63	0.87	-	0.87	0.49
Mexico	1835	0.11	0.73	0.25	0.74	0.58	0.74	-	0.74	0.55
Malaysia	621	0.13	0.83	0.00	0.84	0.47	0.84	-	0.84	0.61
Netherlands	462	0.22	0.75	0.50	0.77	0.56	0.75	-	0.75	0.58
Peru	537	0.11	0.88	-	0.88	0.53	0.88	-	0.88	0.59
Philippines	1016	0.27	0.87	-	0.87	0.59	0.87	-	0.87	0.55
Poland	788	0.27	0.80	0.50	0.82	0.57	0.80	-	0.80	0.60
Portugal	3110	0.35	0.89	-	0.89	0.42	0.89	-	0.89	0.46
Qatar	597	0.34	0.72	0.00	0.72	0.55	0.72	-	0.72	0.52
Romania	496	0.30	0.71	0.00	0.71	0.47	0.71	-	0.71	0.44
Russia	2089	0.11	0.67	0.71	0.67	0.57	0.66	-	0.66	0.57
Sweden	5461	0.62	0.61	0.61	0.56	0.52	0.60	0.60	-	0.52
Slovakia	3486	0.13	0.66	0.40	0.67	0.50	0.67	-	0.67	0.49
Turkey	2773	0.25	0.72	-	0.72	0.52	0.72	-	0.72	0.52
Ukraine	1367	0.21	0.79	-	0.79	0.52	0.79	-	0.79	0.55

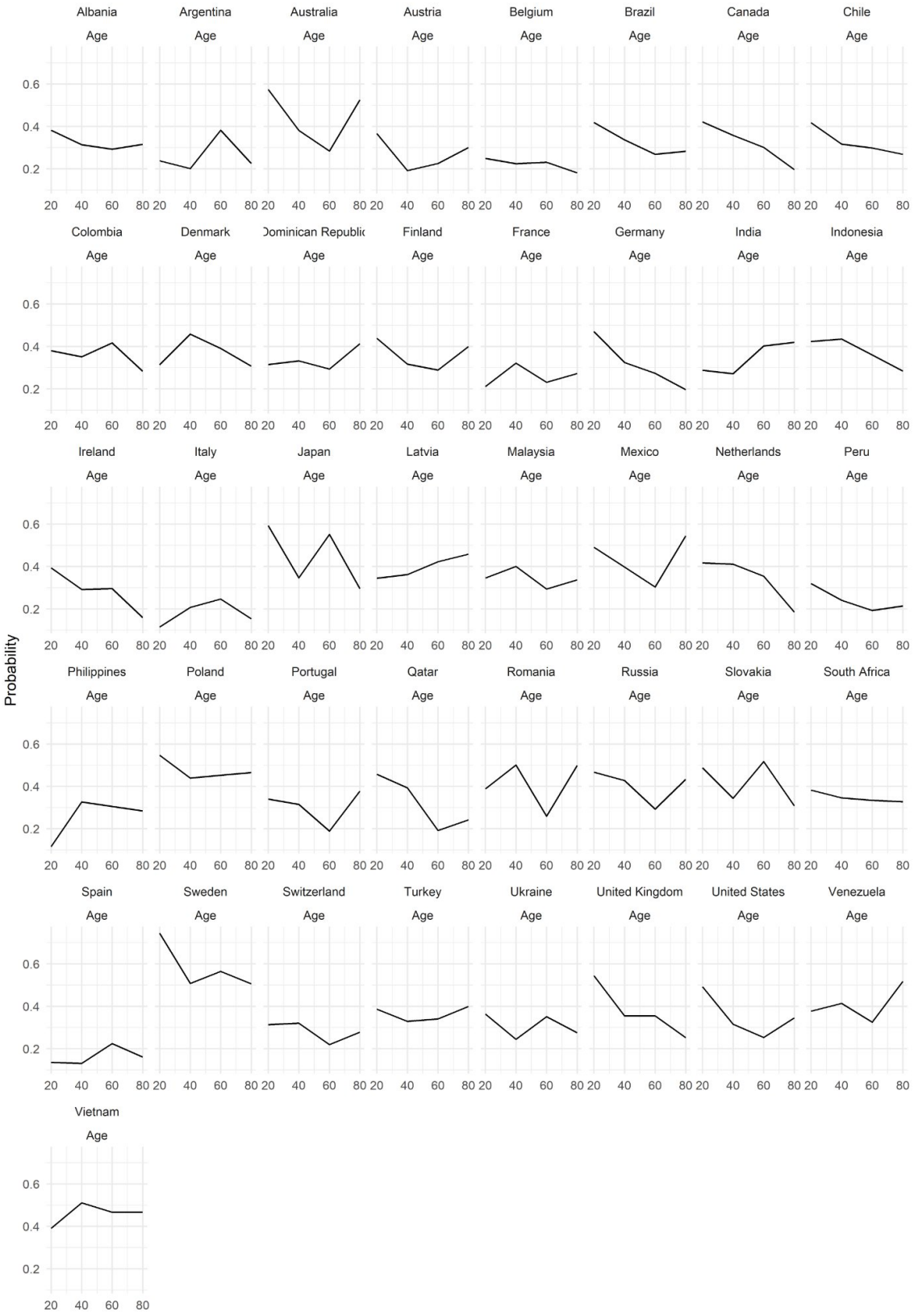
United States	10550	0.26	0.82	0.00	0.82	0.57	0.82	-	0.82	0.59
Venezuela	10686	0.18	0.81	-	0.81	0.48	0.81	-	0.81	0.47
Vietnam	626	0.16	0.63	0.33	0.63	0.54	0.62	0.00	0.63	0.55
South Africa	771	0.31	0.69	0.67	0.69	0.54	0.68	-	0.68	0.60

RF = random forest, acc = accuracy, npv = negative predictive value, ppv = positive predictive value, ROC AUC = area under the ROC curve

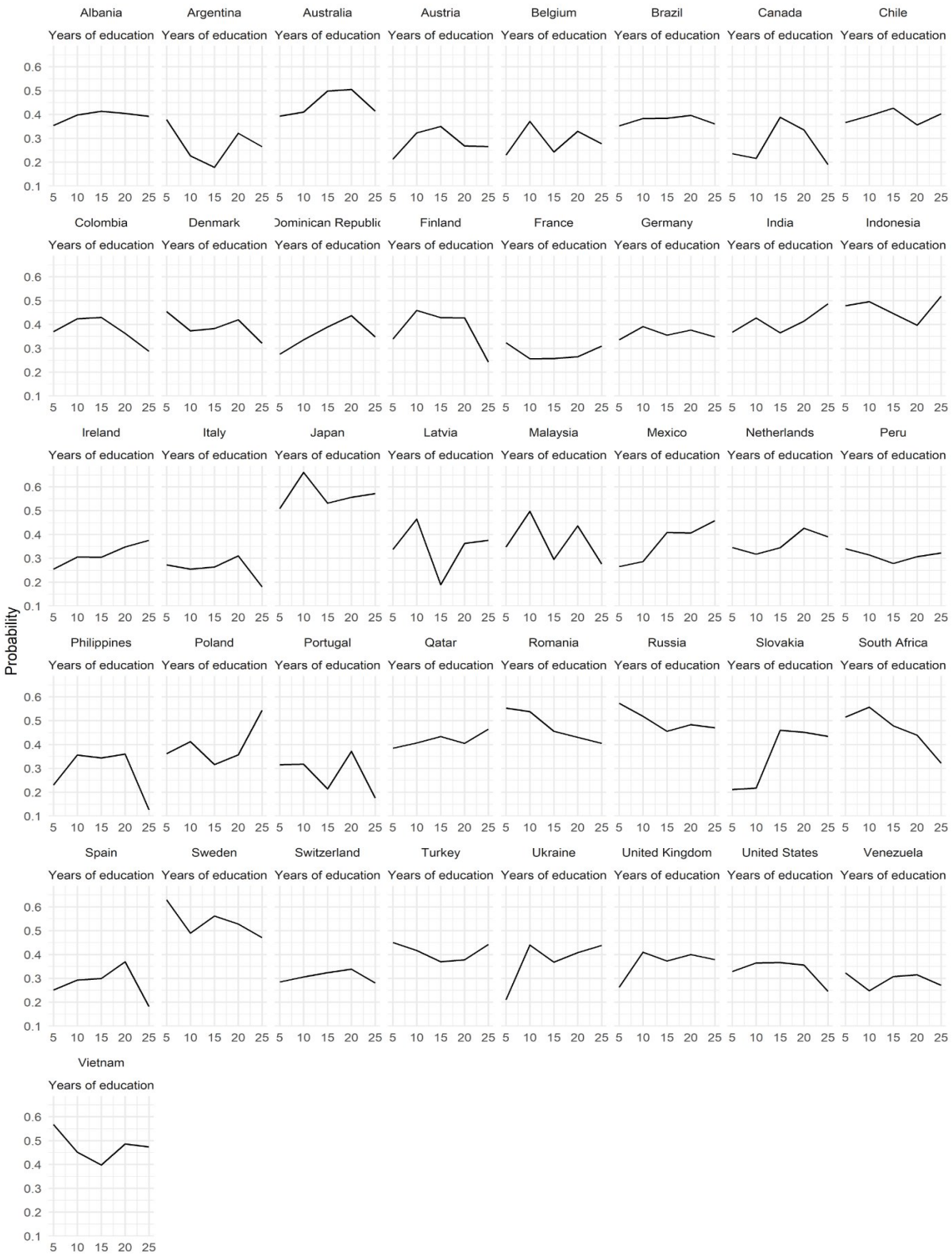
Supplement Figure 1. Accuracy metrics of unweighted random forest and logistic regression models, as a function of sample size.



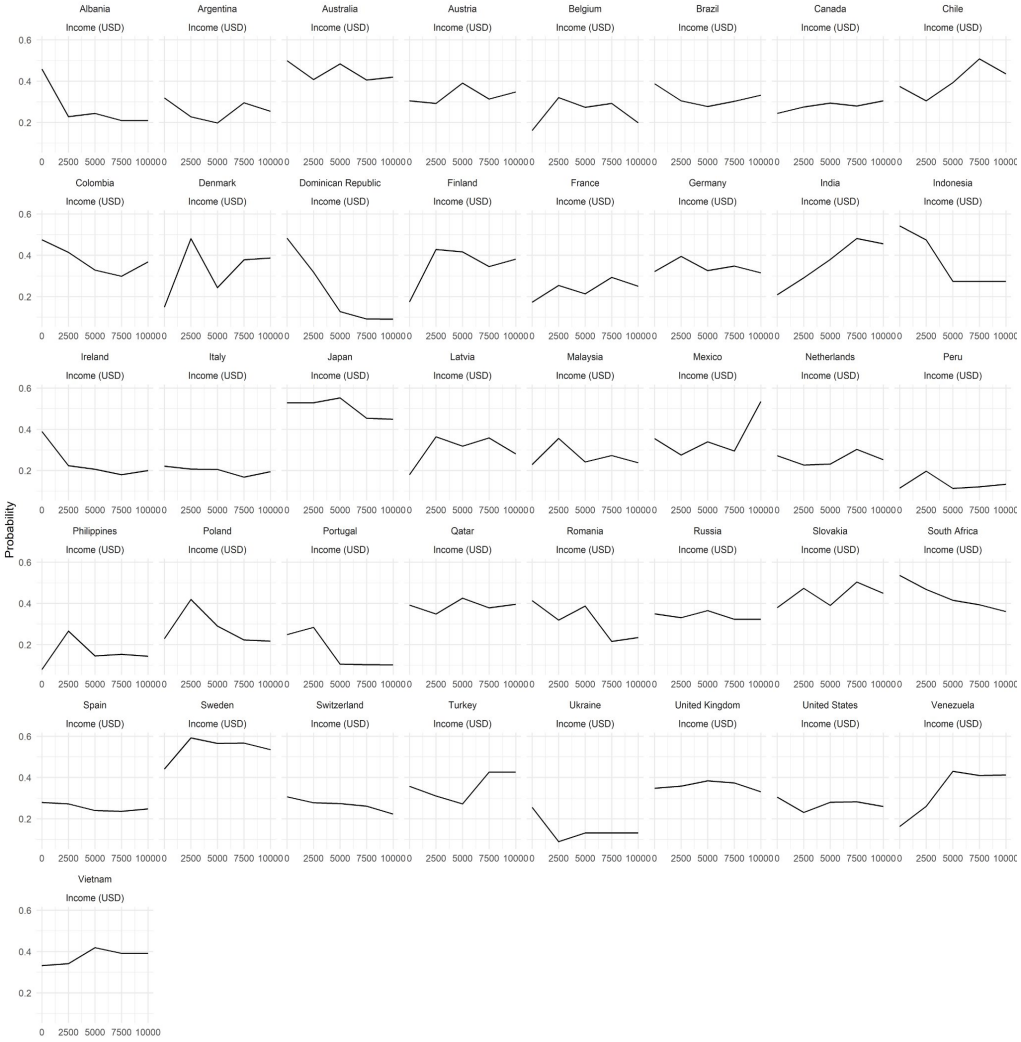
Supplement Figure 2. Partial dependence plots showing the effect of ‘Age’ on attendance of social gathering, by country (weighted model).



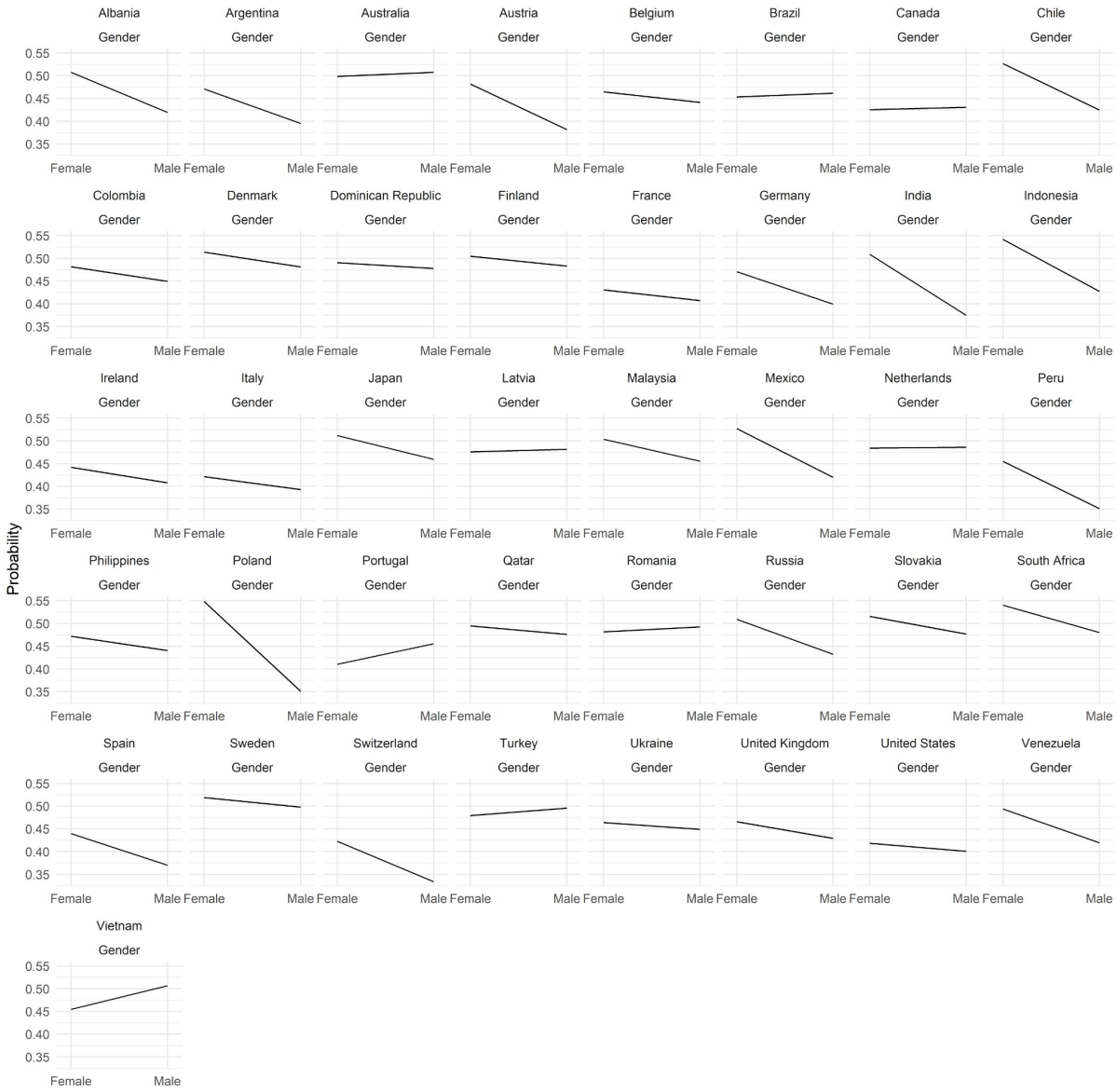
Supplement Figure 3. Partial dependence plots showing the effect of 'Year of education' on attendance of social gathering, by country (weighted model).



Supplement Figure 4. Partial dependence plots showing the effect of 'Income' on attendance of social gathering, by country (weighted model).


Albania
Argentina
Australia
Austria
Belgium
Brazil
Canada
Chile
Colombia
Denmark
Dominican Republic
Finland
France
Germany
India
Indonesia
Ireland
Italy
Japan
Latvia
Malaysia
Mexico
Netherlands
Peru
Philippines
Poland
Portugal
Qatar
Romania
Russia
Slovakia
South Africa
Spain
Sweden
Switzerland
Turkey
Ukraine
United Kingdom
United States
Venezuela
Vietnam
Income (USD)
Probability
0.6
0.4
0.2
0 2500 5000 7500 100000

Supplement Figure 5. Partial dependence plots showing the effect of ‘Gender’ on attendance of social gathering, by country (weighted model).



Discussion

We would like to reaffirm that our sample was not representative, thus the generalizability of our results is smaller. As the upsampling and weighting procedures can introduce biases, we should consider other methods to handle class imbalance. For example, in the next research, different oversampling techniques such as SMOTE, ADASYN, or Borderline-SMOTE could be used. These methods aim to create synthetic samples that are more representative of the rare class, reducing the potential for bias.

As seen on Supplement Figure 1, the unweighted random forest models do not outperform the logistic regression models. As accuracy is somewhat dependent on the base rate, as evidenced by Supplement Figure 1, it is not an informative metric of overall performance. ROC AUC values, on the other hand, serve as an unbiased accuracy metric. We see that these values go over 0.5 most of the time, so the models use information from data to predict better than random; however, the overall accuracies, as evidenced by these ROC AUC values, are still quite low. We can conclude that demographics are only slightly useful in predicting reported social gathering attendance during the COVID-19 pandemic.

Chapter IV.

Szaszi, B., Komandi, K., Hajdu, N., Tipton, E. (2022). Applying behavioral interventions in a new context. In.: Mažar, N., & Soman, D. (Eds.). (2022). *Behavioral Science in the Wild*. University of Toronto Press.

Barnabas Szaszi[1], Krisztian Komandi, Nandor Hajdu[1,2], Elizabeth Tipton[3]

[1] ELTE, Eotvos Lorand University, Hungary

[2] Doctoral School of Psychology, Institute of Psychology, Eotvos Lorand University, Hungary,

[3]Northwestern University, US

Esther worked as a middle manager in the customer service department of a large firm where, as a behavioral science enthusiast, she considered applying nudges. One evening, when reading through the news, she noticed a recent study showing that a small change in communication nudged customers' behavior toward using emails instead of phones when they contacted the customer service department of a multinational corporation. The article discussed that adding the text “87% of our clients in your area prefer to handle their complaints through our website” into the monthly newsletter increased the number of customers using the online form by 17%. As the idea seemed easily applicable and could potentially save thousands of dollars each day for her company, she decided to pitch its implementation to her team the next day. Her boss loved the idea, and in less than one month, an A/B testing was sent out to the target customers. Being curious about the findings, she went to work one hour earlier the day the results came in. However, when she looked at the data, she got very disappointed and confused, wondering what went wrong: it seemed that the intervention had no effect whatsoever on the customers’ behavior. What did Esther do wrong? How could she have minimized the probability of this failure? This puzzling situation is familiar to many who apply behavioral science in the wild. In this chapter, we aim to provide some answers and highlight some rules of thumb and practices to consider when applying behavioral interventions in a new context.

Expect that the effectiveness of nudges vary across contexts

Although this advice seems obvious, when we see the results of experiments backed by data and scientific methods, we tend to believe not just that the results hold but also that they generalize to our specific context [1]. These are often due to some wishful thinking and many books, articles, and keynote presentations where applicability of nudges are presented without drawing attention to their limitations. Nudges, however, often have different effects across different contexts: populations, locations, cultures, and times matter.

Each human being has her own experience, desires, and skills. The same social norm can influence you and your 75 years old grandmother highly dissimilarly. While maybe you could be nudged by social norms to switch to using emails instead of the phone when making complaints, the same message could have no effect at all on your grandmother for countless potential reasons, such as her different perceptions about the norms in her

local network or her lack of familiarity with smartphones or computers. Even the same individuals behave differently across different situations and times. Maybe in the morning you are too stressed to process any new information and you miss the newsletter but you are prone to open your non-work related emails during commuting home after work.

Consider the following widely used example of social norms and nudging. In a study, researchers aimed to reduce energy consumption of US households by providing descriptive information on social norms. In a series of randomized controlled experiments involving roughly 588,000 households (similarly to A/B studies in the online world), an energy management company (Opower) tested whether providing information about the neighbour's consumption, that is, descriptive norms, impacts energy usage [2]. It was estimated that on average, the program decreased energy consumption by 2%, an equivalent effect to an 11-20% electricity price increase. Given that the effects were robust and tested on a huge sample, it was thought that the success could be easily reproduced when scaling up the study to new states and other households. However, in later evaluations, the interventions had practically no effect on energy consumption [3,4]. Although at first sight such failure might be shocking, the careful consideration of the context can provide some answers. While both the context and the details of the nudge remained the same, the types of households included in the scaled up evaluations were different from the original studies and as a result their behavior was much harder to change. Later research revealed that less environmentally friendly attitudes, lower-income and thus smaller households, and beliefs about local support of the provided descriptive norm might have all mitigated the effect of this nudge. Therefore, even in a case where there is seemingly little difference between the original and the new settings, there can be important contextual factors which significantly influence, and even diminish the effectiveness of an intervention. When applying nudges, similarly to Esther or the managers at Opower, we usually have a benchmark example in mind where a behavioral intervention successfully triggered the desired effect in the past. Sometimes, it is reasonable to assume that the nudge will work in our context without deeper consideration but in most of the cases, we need to put considerable energy into figuring out what can go wrong and how the new context is different from the benchmark example we have in mind.

In one of our studies, we aimed to explore all the main factors that influence people's choice in a given situation - in our case it was choosing between the stairs and the elevator[5]. Only using information on the context of the choice, we accurately predicted in 93.26% of the cases whether one chose the stairs or elevator. Although in this study we haven't assessed how the context influenced the effectiveness of an intervention, it is reasonable to assume that if the contextual factors had such an influence on people's choice, they can have similarly influential effects on the effectiveness of behavioral interventions influencing those choices.

The understanding of the context is often the difference between the application of successful and failed nudges[6]. In a recent study using a clever approach, researchers showed that nudges run by two of the largest Nudge Units in the United States on average only had a 1.4% increase on the desired outcomes, an effect much smaller than one would expect when reading the published studies and articles on the topic[7]. (This discrepancy stems in part from the fact that successful studies are published and we hear about them more often, while the failures are frequently buried in the file drawer.) The average effect is relatively small, but it is clearly different from zero (nudges work!), and the list of nudges show huge variance in their effectiveness - ranging from backfiring interventions to highly successful ones.

Explore the contextual factors that may influence the effectiveness of your intervention

It is useful to think about the process of context exploration as analogous to an anamnesis in a therapy setting. No therapist would start a therapy without trying to understand the specific context of the patient. Behavioral interventions should not be employed without thoughtful exploration of the contextual factors either. It can designate the directions of thinking and help you decide about the proper intervention.

The influencing contextual factors can be of many kinds: physical attributes of the environment, nonphysical factors such as social, cultural, or psychological attributes and preferences of the target population, as well as the timing of the choice you want to influence. Another type of influencing factor concerns the behavioral intervention itself: are there specific situations when the intervention is not supposed to work? While there

is no easy way to find all the factors, our emphasis is more on the need for a structured way to understand the context of a given decision or behaviour. In recent years, human centred design (HCD) has become a popular inspiration for organisations that want to better explore and understand their target groups. Complemented with analysing - ideally - behavioural data, HCD might help a decision maker to explore how each of the important dimensions listed above can have an influence. In fact, the combination of behavioural science and HCD has resulted in the new, increasingly popular field of behavioral design[8]. Although this process is not a safeguard for success, it can definitely help identify the biggest holes in the plan. More resources coming from reviewing the literature, reading about similar interventions, as well as conducting interviews and qualitative surveys or even focus group discussions can lead to a more thorough list of the influencing factors.

In the example of Esther, a range of factors can play a role: age, socioeconomic status, place of living, general attitude towards computers or emailing or perceived difficulty of use. Esther could conclude that her listed contextual factors can be compressed into two main categories that she thinks to matter: age and tech savviness. Those who are older, and not very tech savvy won't contact customer service by email whatever nudges she uses, while with the younger and more tech savvy people she will have a much better chance. As she also has information about the customers of her company, this information, and the estimated cost of the intervention could help her decide whether it's worth it to try applying the nudge in her context.

Test the effect of the nudge with contextual diversity in mind

While thinking through the contextual factors can be useful, you need to test the effectiveness of your intervention in a small sample in a context similar to yours if you want to minimize the probability that the scaled-up intervention fails in your context. As there are many books detailing the advantages of randomized testing and describing the ways how to do it, we are not discussing them here. Instead, we will focus on one important and again often ignored attribute of testing, which can define the success of the whole endeavour: testing on a diverse sample.

When you test, using the insights from the context exploration, you need to try to anticipate how the effect of the nudge might vary across your population[9]. Remember that an intervention may be more effective with some than others. For example, in Esther's firm, if you test your intervention on customer groups that typically already use both email and phone, your intervention may work more effectively than on a population that typically only uses phone. Once you have considered the factors that might influence the effectiveness of the nudge, divide your population into subgroups in which you anticipate the nudge to perform similarly. When conducting a test, focus efforts on recruiting subgroups which you expect to be most similar to the population whose behavior you aim to change. That way you can ensure that your test results will be similar to results of the scaled up behavioral intervention. Recall that this was not the case in the Opower study, which resulted in quite different results in the early study than were found in the full population. If you test the intervention in each of the subgroups, you would be able to identify those parts of the population where the nudge might be the most effective. You should keep in mind that this means you will need a large enough sample size in each of the subgroups to estimate the subgroup average effect. Carefully considering these different purposes in advance, however tedious, will allow you to design a study that answers all of your questions, instead of leaving you, like Esther, puzzled.

Conclusion: stay skeptical until you have proof

So what would we recommend to Esther? Unlike in a lab where the context can be controlled, when applying behavioural insights in the wild, we constantly run into new configurations of the factors that might have an impact both on a target behaviour and on the effectiveness of a nudge. Assuming that the effectiveness of the behavioral intervention will vary and may even not work is possibly the best motivation for any behavioural scientist to keep exploring the context of a behaviour. We should not get carried away with any testing opportunity but instead focus on systematically building a diverse test sample that ensures we will end up being able to tell successful and failed nudges apart. After all, when exploring the wild, preparing for surprises is the best strategy one can follow.

References

1. Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A. & Aczel, B. A Systematic Scoping Review of the Choice Architecture Movement: Toward Understanding When and Why Nudges Work. *Journal of Behavioral Decision Making* **31**, 355–366 (2018).

2. Allcott, H. & Rogers, T. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review* **104**, 3003–37 (2014).

3. Allcott, H. Site selection bias in program evaluation. *The Quarterly Journal of Economics* **130**, 1117–1165 (2015).

4. Jachimowicz, J. M., Hauser, O. P., O'Brien, J. D., Sherman, E. & Galinsky, A. D. The critical role of second-order normative beliefs in predicting energy conservation. *Nature Human Behaviour* **2**, 757–764 (2018).

5. Hajdu, N., Szaszi, B. & Aczel, B. Extending the Choice Architecture Toolbox: The Choice Context Mapping. (2020).

6. Tipton, E., Bryan, C. J. & Yeager, D. S. To change the world, behavioral intervention research will need to get serious about heterogeneity. *Manuscript in Preparation. Retrieved from https://statmodeling. stat. columbia. edu/wp-content/uploads/2020/07/Heterogeneity-1-23-20-NHB. pdf* (2020).

7. DellaVigna, S. & Linos, E. *Rcts to scale: Comprehensive evidence from two nudge units*. (2020).

8. Tantia, P. The new science of designing for humans. *Stanford Social Innovation Review* **15**, 29–33 (2017).

9. Tipton, E., Yeager, D. S., Iachan, R. & Schneider, B. Designing probability samples to study treatment effect heterogeneity. *Experimental methods in survey*

research: Techniques that combine random sampling with random assignment 435–456 (2019).

Discussion

The incorporation of machine learning methodologies into exploratory research within psychology and related fields presents a significant opportunity to reshape our understanding of human behavior. These approaches hold the potential to serve as a catalyst for expanding the boundaries of knowledge in the domain of psychology. In this dissertation, three research papers contribute empirical findings and methodological insights to the field of behavioral science, focusing on the intricate relationship between contextual factors and human choices. The first paper introduces the Choice Context Exploration, a systematic approach aimed at comprehensively understanding the contextual determinants of choice behavior. This method elucidates influential factors and underlying belief structures, offering a structured framework for tailoring future interventions. The second paper examines global adherence to pandemic lockdown measures, revealing the consistent significance of fear, responsibility, and social influences while emphasizing the necessity for region-specific strategies. The third paper investigates the association between demographic variables and social gathering behavior during epidemics across 41 countries, highlighting the complex interplay between context and individual choices. Lastly, a book chapter underscores the context-dependent nature of behavioral interventions, advocating for a cautious and adaptable approach, grounded in empirical testing and skepticism. Collectively, these studies contribute to the growing body of knowledge on the role of contextual factors in shaping human behavior and provide practical implications for the development of more effective, context-sensitive interventions. The incorporation of machine learning methodologies into exploratory research within psychology and related fields not only presents a significant opportunity to reshape our understanding of human behavior but also underscores the importance of handling complexity in behavioral science. These approaches offer a nuanced perspective on the intricate interplay between contextual factors and human choices, highlighting the need for sophisticated analytical techniques to unravel the complexities inherent in behavioral phenomena. After the concise summarization of the four articles presented, I discuss prospective avenues of further research.

Synopsis of Chapters I-IV

Chapter I of this dissertation centers on the challenges faced by practitioners of choice architecture interventions when adapting such interventions to new contexts. The paper argues for an approach that begins with a comprehensive examination of contextual factors influencing the targeted choice. We introduce the Choice Context Exploration, a three-step procedure showcased in university students' choices between stairs and elevators.

In Step 1, 15 potential contextual factors were collected, and in Step 2, based on our survey, we estimated their effects on participants' behavior. Factors such as peer choices, destination floor, environmental consciousness, and health aspirations emerged as influential. The procedure achieved over 90% accuracy in predicting choices based on contextual information. Step 3 identified three distinct belief groups among participants: "Comfort-driven," "Principles-driven," and "No priority." People correctly assessed the importance of the destination floor but their beliefs predict their reported choice less accurately than the reported contextual factors.

The main benefit of using the Choice Context Exploration is that it is a systematic and easy to follow way of identifying influencing factors and beliefs, that might facilitate intervention planning, and tailoring interventions to specific belief groups. It's especially useful for new choice situations, environments, or populations, reducing costs and minimizing the risk of counterproductive interventions. Limitations included the potential for new factors to emerge, changes in factor effects over time, and the need to explore interactions. Sample size and generalizability concerns were noted, and self-reported measurements were used. Future research should explore different subpopulations and additional aspects in intervention design.

In summary, the Choice Context Exploration offers a systematic approach to understanding contextual factors, enhancing intervention planning, and possibly increasing the effectiveness of choice architecture interventions. The Choice Context Exploration exemplifies a systematic approach aimed at disentangling the multifaceted influences shaping choice behavior. By identifying and analyzing a multitude of contextual factors, this method offers insights about decision-making, while addressing the need to navigate complexity in intervention planning. Further research can expand its

application to diverse environments and examine moderating factors in implementing nudges.

Chapter II of this dissertation delved into the significance of contextual factors in predicting individuals' decisions to adhere to pandemic lockdown measures. These factors demonstrated consistent patterns across 16 different countries, highlighting the robustness of the findings across diverse settings. Boredom and fellow citizens' non-compliance with regulations consistently increased the likelihood of leaving home, while the fear of infection and a sense of responsibility consistently decreased it.

While specific factors varied in importance among countries, some key trends emerged. The *fear of getting infected* ranked high in 12 countries, with heightened effects in Hungary, Japan, the Netherlands, Romania, Switzerland, and the UK, and less impact in Greece, Nigeria, and the Russian Federation. The *feeling of responsibility* toward society emerged as a top-three factor in 11 countries. However, this relationship was weaker in Greece, Japan, and Switzerland, where the *adherence of fellow countryfolk* played a more significant role. The *feeling of being caged while at home* was influential in eight countries, most notably in the UK and Slovakia. However, it had little impact in Japan and Greece.

One of the primary limitations of this research lies in the composition of the study samples. The participants were drawn from specific countries, and the data collection methods, rates of infection, and lockdown recommendations varied within and between these countries. This limited scope raises concerns about the generalizability of the findings to a broader global population. Moreover, the study largely focused on developed and developing countries, potentially neglecting insights from underrepresented regions with vastly different COVID-19 countermeasures. The research depended on risk score predictions to understand adherence to lockdown recommendations. However, the prediction accuracy was notably low for certain countries. Additionally, the reliance on participants' recollection of their situations, especially in countries not under lockdown during data collection, introduces recall bias and potentially undermines the accuracy of risk score predictions.The sample may not fully mirror the broader population, potentially affecting the generalizability of the findings.There were differences between countries in terms of which factors were most critical in predicting compliance with stay-at-home orders. This variability underscores the complexity of human behavior and suggests that

one-size-fits-all interventions may not be effective. However, it also presents a challenge in creating cohesive public health strategies.

In summary, the *fear of getting infected* emerged as the most critical predictor of adherence to lockdown measures, alongside feelings of *responsibility*, *being caged*, and the perceived *compliance of fellow citizens*. These findings might have public health implications, indicating that it could be helpful to focus on personal responsibility and the collective effort in dealing with pandemics. Additionally, there's potential for promoting a more positive view of confinement and addressing social isolation. Moreover, transparent information about disease risks could possibly improve adherence to lockdown regulations. These potential implications, however, should be tested by measuring behavior instead of reported behavior.

In our third paper presented in this dissertation, we analyzed a large dataset collected from 41 countries during the early phase of the COVID-19 pandemic to investigate the association between demographic factors and individuals' tendencies to attend or avoid social gatherings during epidemics. The study covered countries representing 73.05% of the world's population. The findings revealed several general patterns across most countries: a higher proportion of males, younger individuals, lower-educated individuals, and those with lower incomes were more likely to attend social gatherings. However, notable heterogeneity existed among countries, with varying directions of associations between demographic factors and social gathering tendencies. For instance, in some countries, higher-educated citizens were more likely to attend social gatherings than their lower-educated counterparts. These variations emphasize the importance of considering specific contexts and discourage generalizing behaviors from one country to another.

When it comes to targeting interventions based on a single demographic variable, the study suggested that income was the strongest predictor worldwide, followed by age, education, and gender. However, significant differences existed between countries, even among those with similar geographical and cultural backgrounds. The study highlighted that contextual factors strongly influence the relationship between demographic factors and social gathering behavior, emphasizing the need for country-level exploration.

Despite small effect sizes, we emphasized the meaningful consequences of these associations, with even slight differences in social gathering behavior potentially leading

to significant impacts on the spread of diseases. Targeting less avoidant demographic subgroups in public health campaigns could save resources and tailor messages for greater effectiveness. We also pointed out the importance of investigating moderating factors behind these patterns, including differences in government policies, perceived self-efficacy, cultural variations, trust in authorities, and job types.

Similar to Chapter II, in the study presented in Chapter III we collected data during the early phase of the COVID-19 pandemic. The rapid evolution of the pandemic and public perceptions over time may limit the generalizability of findings to later stages of the crisis. The research relied on self-reported data, which can introduce biases and inaccuracies due to memory recall and social desirability. While the study found self-reported data to align with actual social distancing behavior, the potential for measurement error exists. The findings regarding demographic factors associated with social gathering behavior may not necessarily generalize to other pandemics with different characteristics and risks. Factors driving decisions during one pandemic may not hold true for future public health crises. We recognized that countries implemented diverse policies and regulations during the pandemic, which could influence social gathering behavior. However, we did not comprehensively analyze the impact of these policies on individual choices.

In conclusion, this research provided insights into the complex relationship between demographic factors and social gathering behavior during epidemics. It highlighted the importance of context-specific approaches in public health interventions and the need for further research to explore causal explanations for these patterns.

Behavioral interventions, often facilitated through nudges, have gained prominence for their potential to influence human behavior. These subtle changes in choice architecture can yield impressive results, but their effectiveness is highly context-dependent, as exemplified by a study aiming to reduce energy consumption in US households. In Chapter IV, we explored the nuances of behavioral interventions and offered insights into navigating their complexities.

Behavioral nudges, while powerful, are not universally applicable. Their impact can differ significantly across contexts, populations, and scenarios. The assumption that what works in one context will replicate elsewhere is a common pitfall. Factors such as

individual experiences, preferences, and skills influence how individuals respond to nudges. Context exploration is akin to conducting an in-depth assessment before therapy. Similarly, behavioral interventions should be approached with a profound understanding of contextual factors. These factors can encompass physical attributes, social, cultural, or psychological aspects, and even the timing of the targeted behavior. Behavioral interventions themselves may have limitations based on context.

Human-centered design (HCD) principles, coupled with behavioral science and data analysis, form the foundation of behavioral design—a field dedicated to understanding and leveraging context. While context exploration is vital, additional resources such as literature reviews, interviews, and qualitative surveys can enhance the identification of contextual factors.

Testing behavioral interventions should consider the diversity of the target population. Anticipating how the intervention might vary across subgroups is crucial. Subgroup-specific testing provides a nuanced understanding of the intervention's impact, aligning test results with real-world outcomes. For instance, encouraging customers to use emails instead of phone calls may yield varying results based on age or tech-savviness. Conducting tests within these subgroups ensures that results reflect the expected outcomes in the scaled-up intervention. This approach mitigates disparities between test and real-world outcomes.

In the realm of behavioral science, skepticism is a valuable asset. It acknowledges that behavioral interventions are intrinsically tied to context. Expecting variability and potential failure is a powerful motivator for practitioners to diligently explore context. Practitioners should refrain from assuming that a successful nudge in one context guarantees success elsewhere. Instead, they should commit to understanding context, conducting diverse tests, and maintaining skepticism until empirical evidence supports their interventions. This approach ensures that behavioral science remains a dynamic field capable of adapting to the intricacies of real-world contexts. All in all, the discussion on behavioral interventions in the book chapter emphasizes the inherent complexity of influencing human behavior. While behavioral nudges offer promising avenues for intervention, their effectiveness is highly context-dependent, necessitating a nuanced understanding of contextual factors and careful consideration of individual differences.

What have we achieved by incorporating machine learning techniques and algorithms into this research? The practice of splitting our data into training and test sets has proven to be effective in mitigating overfitting, as demonstrated in Chapter 1. Splitting data into training and test sets is a common practice in machine learning and statistical modeling. It involves dividing a dataset into two distinct subsets for the purpose of developing and evaluating a predictive model. The training set is a portion of the dataset that is used to train and build the predictive model. It contains a significant portion of the data and serves as the foundation for the model to learn patterns, relationships, and trends within the data. The model is fitted to this training set, which means it adjusts its parameters and algorithms to capture the underlying characteristics of the data. The test set is a separate portion of the dataset that is held back and not used during the training phase. Instead, it is used to assess the performance of the model after it has been trained. The test set is used to simulate the real-world application of the model, allowing you to evaluate how well the model generalizes from the training data to make predictions on new, unseen data.

The primary purpose of splitting data into training and test sets is to evaluate the model's ability to make accurate predictions on data it has not been exposed to during training. It helps identify whether the model has learned to recognize genuine patterns in the data or if it has overfit the training data (meaning it has memorized the training data but cannot generalize to new data). This process is critical for assessing the model's performance, making adjustments if necessary, and ensuring that it is reliable for making predictions in real-world scenarios.

The accuracy of a model, regardless of the specific accuracy metric employed, consistently appears higher on a training set than on a test set. By partitioning our data in this manner, we have prevented an overestimation of accuracy. However, it is worth noting that in Chapter 3, the random forest algorithm did not surpass the performance of the logistic regression models. This suggests that non-linear effects, which the random forest algorithm is designed to detect but the logistic regression algorithm may miss, were not present in our dataset. The value of the random forest algorithm becomes particularly evident when dealing with datasets that exhibit non-linear relationships between variables and a substantial number of predictors. In our specific case, neither of these conditions applied. Random forests can, at times, exhibit overfitting, as observed in certain models in Chapter 2. Nevertheless, the instances where models clearly overfitted, as evidenced

by their lower accuracy compared to the base rate, were relatively few. All in all, our research indicates that machine learning tools can be useful and informative when used during exploratory analyses. However, these methods are not extensively used in the analysis of psychological experiments as compared to other fields, for example genetics (Orrù et al., 2020). What are the possible reasons behind this?

Why is the use of machine learning tools in psychology research not widespread?

One of the primary barriers to the adoption of machine learning in psychology is the limited awareness and understanding of these techniques among psychologists. Machine learning involves complex algorithms and statistical methods that may not be familiar to researchers in the field. Many psychologists are trained in traditional statistical approaches and may be apprehensive about delving into the complexities of machine learning. Machine learning requires specialized knowledge and expertise in areas such as data preprocessing, feature engineering, model selection, and hyperparameter tuning. Most psychology programs do not offer comprehensive training in these areas, leaving researchers ill-equipped to apply machine learning techniques effectively. As a result, there is a shortage of psychologists with the necessary skills to harness the power of machine learning. Education and training initiatives must evolve to meet the growing demand for expertise in leveraging machine learning in exploratory research. Integrating machine learning courses and workshops into psychology and social science curricula can bridge the gap between domain knowledge and computational proficiency, empowering researchers to harness these technologies effectively. This is the reason why some of the authors of the presented studies collaborated again to write a guide to doing exploratory analysis with machine learning tools, for Hungarian audiences (Hajdu et al., 2023). Another potential solution could involve promoting greater interdisciplinary collaboration, especially with data scientists.

In psychology research, access to large, high-quality datasets can be limited. Integrating machine learning with traditional psychological research methods can be challenging. Researchers may be uncertain about how to combine machine learning techniques with established theories and methodologies. Bridging the gap between these two approaches requires careful planning and interdisciplinary collaboration. Many psychological studies

rely on small sample sizes, which may not be suitable for training complex machine learning models. Moreover, ensuring data quality and integrity is challenging in psychology, as self-reported or survey-based data may be subject to biases and inaccuracies.

Further directions

All in all, we gathered important insights from the previously described research. We are convinced that machine learning techniques are invaluable in doing informative exploratory analyses. Although we used these methods in a context of choice context exploration, they can be applied to any exploratory research in psychology, where measurements are available. We also showed that while insights can be gathered even from only demographic data, the prediction accuracy, and thus, the potential contribution to theory formation is increased when there is more available data. Moving forward, addressing complexity in behavioral science requires a multidisciplinary approach that integrates machine learning techniques with traditional psychological research methods. Prioritizing the collection of diverse datasets and exploring advanced analytical techniques can provide deeper insights into the complex dynamics of human behavior.

Data diversity

First and foremost, future endeavors should prioritize the collection of more comprehensive and diverse datasets. Expanding the range of variables to encompass psychological traits, environmental factors, and behavioral indicators could provide a more holistic understanding of human behavior. In parallel, using feature engineering techniques is crucial to create more informative variables. Approaches such as feature selection and extraction methods can uncover latent patterns in data that may remain concealed when only considering raw measurements. As the availability of data from various sources continues to grow, the integration of multimodal data presents an exciting frontier. Combining textual, numerical, image, audio, and sensor data can provide a richer understanding of complex phenomena. For instance, in psychology research, incorporating facial expressions, voice modulation, and physiological signals alongside traditional survey data can offer deeper insights into human behavior (Akkus et al., 2023).

Different analyses

Another avenue to consider is to employ different analyses in further research. For example, contextual analysis, complemented by Natural Language Processing (NLP) techniques for analyzing text data could provide valuable contextual information that enriches our understanding of decision-making processes. For example, NLP can be used to aid decision-making in healthcare by processing a large number of electrical health records (Hossain et al., 2023). Temporal analysis and longitudinal studies represent a compelling avenue for exploration. The dynamic nature of human behavior calls for investigations that extend beyond cross-sectional data. Leveraging time-series analysis techniques and conducting longitudinal studies can offer insights into how behavior evolves over time, particularly in response to external events like pandemics. One study examined the emotional impact of the COVID-19 pandemic in the UK through natural language processing of Twitter data from news channels and user comments (Evans et al., 2023). They found sadness as the prevalent emotion in news tweets declining over time while anger holds sway among user tweets. Integrating multimodal data presents an exciting opportunity. Human behavior is often influenced by a multitude of factors, including sensory inputs and physiological responses. Combining data from various modalities, such as physiological sensors, eye-tracking devices, and audio recordings, can provide a more comprehensive view of behavior. Advanced machine learning methods, including multimodal fusion techniques and deep learning models, can be harnessed to analyze and model complex interactions among different data streams (Jabeen et al., 2023). One study proposes a multimodal deep-learning framework for early diagnosis of Alzheimer's disease, achieving high accuracy by analyzing longitudinal MRI volumes and cross-sectional biomarkers, with an added explainability module aiding domain experts in understanding diagnostic outputs (Rahim et al., 2023). Another study (Xie et al., 2024) utilizes deep-learning models to identify key features in anti-vaping Instagram image posts associated with high user engagement, highlighting the significance of educational warnings and health risk captions in communicating the dangers of vaping on social media, especially among youth and young adults.

In scenarios where labeled data is scarce or expensive to obtain, semi-supervised and active learning techniques hold promise (Bachman et al., 2017; Hady & Schwenker, 2013). These approaches allow models to learn from both labeled and unlabeled data,

reducing the annotation burden. Researchers can actively query the most informative data points for labeling, optimizing resource utilization. Exploratory studies frequently involve sequential decision-making processes. Reinforcement learning (Sutton & Barto, 2018), which excels in such domains, can be applied to optimize interventions, recommendations, or treatments over time. As the scale of data grows, scalability becomes paramount. Cloud-based machine learning platforms provide the necessary infrastructure for handling large datasets and complex models, offering researchers the computational power needed to tackle ambitious exploratory projects.

Establishing causation, rather than merely identifying correlations, remains a fundamental challenge in exploratory research. Machine learning methods can facilitate the identification of causal relationships, such as the impact of specific interventions on behavior. Techniques like counterfactual analysis, causal inference methods (e.g., causal forests or causal inference networks), and randomized controlled trials (RCTs) can help researchers delve deeper into the causal effects of different factors on human behavior.

Explainable AI

Within the context of utilizing machine learning in exploratory psychology research, one emerging area of paramount importance is *Explainable AI* (XAI; Arrieta et al., 2020). This transparency can aid researchers in formulating hypotheses and refining their investigations. XAI represents a crucial frontier, particularly when machine learning is applied to uncover psychological insights in a diverse array of contexts. In the realm of machine learning, especially with the advent of deep learning techniques, a common predicament arises – the "black box" problem. Imagine employing a machine learning model to analyze intricate psychological patterns, only to receive predictions devoid of any comprehensible rationale. This opacity is a significant challenge, as it obstructs researchers from grasping the decision-making mechanisms of these models. Recent EU regulations stipulate that individuals have the right to comprehend the reasoning behind automated decisions that concern them. This implies that fully opaque or "black-box" approaches are prohibited, especially in areas such as law and insurance. This is one of the key objectives of the EU AI Act. In the context of exploratory psychology research, the lack of interpretability can pose several hindrances. The first one is that this lack of interpretability makes hypothesis generation more difficult. The second one is that bias in AI models is a significant concern, especially when investigating human behavior. The

ability to comprehend how and why a model makes specific predictions is vital for detecting and mitigating bias effectively. XAI tries to answer these challenges, as it aims to render AI systems more transparent and comprehensible. XAI techniques offer human-readable explanations, allowing researchers to identify which features or data points influenced specific predictions. This provides valuable context for exploration studies. XAI facilitates the detection of bias within AI models by revealing decision processes. Researchers can discern if certain demographics are treated unfairly, enabling proactive bias mitigation. Examples of XAI techniques include Local Interpretable Model-agnostic Explanations (LIME; Ribeiro et al., 2016): LIME generates locally faithful explanations for individual predictions by fitting simple, interpretable models around smaller groups of data points.

Ethical considerations

As machine learning's role in exploratory research continues to expand, ethical considerations become increasingly prominent. Researchers must prioritize ethical data collection, ensure transparency in algorithms, and implement bias mitigation techniques to mitigate ethical concerns related to data privacy, bias, and fairness. Exploring fairness-aware machine learning and fairness metrics can address bias and discrimination, rendering research results more equitable and inclusive. Reproducibility and open science practices are essential for the credibility of research findings. Ensuring that datasets, code, and methodologies are shared facilitates the replicability of results and fosters collaboration among researchers. Nevertheless, we must also be cautious about potential ethical issues that could emerge when revealing these assets. A balance needs to be struck, where, depending on the model, at least some of the model parameters are not disclosed to the public. The ethical utilization of these assets should not only be promoted but also ensured to prevent any unethical use. Developing standardized pipelines and tools for exploratory research can streamline the process and make it more accessible across various domains.

Expert-expert, and expert-AI collaborations

Collaboration emerges as another pivotal avenue for exploration. Cross-disciplinary partnerships between machine learning experts, psychologists, and researchers from domain-specific fields hold the potential for more meaningful insights. The synergy

between domain experts, who provide guidance in framing research questions and interpreting results, and machine learning specialists, who develop novel algorithms and models, can amplify the impact of studies. Human-AI collaboration represents a transformative paradigm (Wang et al., 2020). AI systems can assist researchers in generating hypotheses, identifying patterns, and automating repetitive tasks, freeing experts to focus on interpretation and theory building. The development of human-AI collaboration platforms and tools can foster synergistic interactions between researchers and intelligent algorithms.

In conclusion, by embracing the challenge of handling complexity in behavioral science, researchers can unlock new avenues for understanding human behavior and developing more effective interventions to address complex societal challenges. By embracing these diverse directions, researchers can unlock new horizons for discovery, contributing to a more profound understanding of the intricate facets of human decision-making and behavior. This evolution promises to be instrumental in advancing the boundaries of knowledge in these domains.

References

Akkus, C., Chu, L., Djakovic, V., Jauch-Walser, S., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., Schulte, R., Urbanczyk, K., Goschenhofer, J., Heumann, C., Hvingelby, R., Schalk, D., & Aßenmacher, M. (2023). *Multimodal Deep Learning*. https://doi.org/10.48550/ARXIV.2301.04856

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bachman, P., Sordoni, A., & Trischler, A. (2017). Learning algorithms for active learning. *International Conference on Machine Learning*, 301–310.

Evans, S. L., Jones, R., Alkan, E., Sichman, J. S., Haque, A., de Oliveira, F. B. S., & Mougouei, D. (2023). The emotional impact of COVID-19 news reporting: A longitudinal study using natural language processing. *Human Behavior and Emerging Technologies*, *2023*.

Hady, M. F. A., & Schwenker, F. (2013). Semi-supervised Learning. In M. Bianchini, M. Maggini, & L. C. Jain (Eds.), *Handbook on Neural Information Processing* (Vol. 49, pp. 215–239). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36657-4_7

Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., & Pisani, A. R. (2023). Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *arXiv preprint arXiv:2306.12834*.

Jabeen, S., Li, X., Amin, M. S., Bourahla, O., Li, S., & Jabbar, A. (2023). A review on methods and applications in multimodal deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(2s), 1-41.

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in psychology*, *10*, 2970.

Rahim, N., El-Sappagh, S., Ali, S., Muhammad, K., Del Ser, J., & Abuhmed, T. (2023). Prediction of Alzheimer's progression based on multimodal Deep-Learning-based fusion and visual Explainability of time-series data. *Information Fusion*, *92*, 363-388.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020, April). From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1-6).

Xie, Z., Deng, S., Liu, P., Lou, X., Xu, C., & Li, D. (2024). Characterizing anti-vaping posts for effective communication on Instagram using multimodal deep learning. *Nicotine and Tobacco Research*, *26*(Supplement_1), S43-S48.