Promoting Data Science

Volume 2 2022 Number 1

Journal of Behavioral Data Science V2N1 (2022)

https://isdsa.org

JOURNAL OF BEHAVIORAL DATA SCIENCE

ISBN: 2575-8306 (Print) 2574-1284 (Online)

https://jbds.isdsa.org

JOURNAL OF BEHAVIORAL DATA SCIENCE

No Publication Charge and Open Access

jbds@isdsa.org

List of Articles

Journal of Behavioral Data Science, 2022, *2* (1), 1–34.
DOI: https://doi.org/10.35566/jbds/v2n1/p1

The Lighting of the BECONs: A Behavioral Data Science Approach to Tracking Interventions in COVID-19 Research

Denny Borsboom*[1], Tessa F. Blanken[1], Fabian Dablander[1], Frenk van Harreveld[1], Charlotte C. Tanis[1], and Piet Van Mieghem[2]

[1] Department of Psychological Methods, University of Amsterdam
d.borsboom@uva.nl
[2] Delft University of Technology

Abstract. The imposition of lockdowns in response to the COVID-19 outbreak has underscored the importance of human behavior in mitigating virus transmission. The scientific study of interventions designed to change behavior (e.g., to promote physical distancing) requires measures of effectiveness that are fast, that can be assessed through experiments, and that can be investigated without actual virus transmission. This paper presents a methodological approach designed to deliver such indicators. We show how behavioral data, obtainable through wearable assessment devices or camera footage, can be used to assess the effect of interventions in experimental research; in addition, the approach can be extended to longitudinal data involving contact tracing apps. Our methodology operates by constructing a contact network: a representation that encodes which individuals have been in physical proximity long enough to transmit the virus. Because behavioral interventions alter the contact network, a comparison of contact networks before and after the intervention can provide information on the effectiveness of the intervention. We coin indicators based on this idea Behavioral Contact Network (BECON) indicators. We examine the performance of three indicators: the Density BECON, the Spectral BECON, and the average shortest path length (ASPL) BECON. First, the Density BECON is based on differences in network density, i.e., differences in the portion of realized edges (connections) relative to all potential edges. Second, the Spectral BECON is based on differences in the eigenspectrum of the adjacency matrix, which capture the spreading potential of the virus. Third, the ASPL BECON is based on differences in the mean of all the shortest distances (i.e., number of edges) between each pair of nodes in the network, and captures the average distance between nodes. Using simulations, we show that all three indicators can effectively track the effect of behavioral interventions. Even in conditions with significant amounts of noise, BECON indicators can reliably identify and order effect sizes of interventions. The present paper invites further study of the method as well

as practical implementations to test the validity of BECON indicators in real data.

Keywords: Contact networks · Network analysis · Epidemiology · Virus spread · Interventions

1 Introduction

The COVID-19 outbreak has underscored the importance of human behavior in controlling virus transmission. As long as vaccines are not operational, the only way to influence transmission rates is through behavioral interventions that either prohibit specific kinds of behavior (e.g., attending school, visiting relatives, leaving the house) or promote others (e.g., physical distancing, wearing masks, complying with regulations). As such, behavior is fundamental to important parameters in epidemiological models, such as the reproduction number (the number of people a randomly chosen disease carrier is expected to infect): even though virus transmission depends on biological characteristics of the virus and the human system, its speed reflects an interaction between biology and behavior (Delamater, Street, Leslie, Yang, & Jacobsen, 2019; Heesterbeek et al., 2015). Indeed, one way of understanding the reasoning behind lockdowns is that they try to drive down the reproduction number by changing behavioral patterns (de Vlas & Coffeng, 2021; Jeffrey et al., 2020). The goal of this paper is to contribute to our understanding of these behavioral patterns, by developing methodological tools that can be used to study them.

To successfully monitor and control our responses to a virus outbreak like COVID-19, we need to obtain insight into the relative effectiveness of different behavioral interventions. Relevant behavioral interventions can either be implemented at a microlevel (e.g., setting up nudges in a store to promote physical distancing, changing the floor plan of a restaurant), or a macrolevel (e.g., implementing public policy measures that promote working from home, closing public buildings). Currently, however, methodology for estimating effects of such interventions at the behavioral level is limited to highly indirect assessments based on measures of virus spread. For example, comprehensive assessments of interventions at the macrolevel (Chu et al., 2020) have been estimated based on the relation between country-level interventions (e.g., school closings, lockdowns) and corresponding population statistics (e.g., hospital admissions, IC uptakes, death rates; see for example Flaxman et al., 2020); or they have been treated as model parameters to assess the time-course of the epidemic under different scenarios – the well-known study by Ferguson et al. (2020), which has played an important role in COVID-19-related policy, is a case in point.

There are at least three methodological reasons why indicators such as hospital admissions are of limited use in assessing effects of behavioral interventions designed to counter virus spread. The first problem is that they are *lagged indicators*. Evaluating the effect of interventions with hospital admissions as a dependent variable suffers from the time course of virus transmission, incubation, and disease progression, before one can assess where the intervention has

been effective (this delay was two to three weeks for COVID-19). The second problem concerns *experimental inaccessibility*. If one studies an intervention that is strongly suspected to be effective, it is unethical to install a control group for comparison and to wait for participants to become ill. The next best alternative — a quasi-experimental research setup — suffers from considerable levels of confounding, and because interventions are almost always implemented in packages it is hard to disentangle their effects. Third, current indicators require an *active virus*. Thus, in a period in which there is no virus active, it is impossible to study the effects of behavioral interventions. This is strategically impractical as it would be ideal to study behavioral interventions while the virus is inactive in order to prepare for a possible future outbreak. Moreover, the COVID-19 pandemic is unlikely to be the last global pandemic and research in effective interventions will remain important, even after the current crisis has ended.

The scientific study of behavioral interventions thus requires indicators that are fast, that can be assessed through experiments, and that can be investigated without actual transmission of the virus. In the present paper, we develop a methodological approach designed to deliver such indicators. In a nutshell, we make use of the fact that behavioral data, obtainable through wearable devices, camera data, or tracing apps can be used to assess contact networks (Cencetti et al., 2020). This can either be done at the microlevel (e.g., assessing contact patterns at a public gathering) or at the macrolevel (e.g., reconstructing contact networks at the level of a city on the basis of tracing apps). We coin indicators based on such networks Behavioral Contact Network (BECON) indicators. Because BECON indicators are available in real time, they respond to induced changes in contact networks virtually instantaneously; and because they do not require actual transmission of the virus, they can be used to assess effectiveness in healthy subjects, which in turn means they can be studied in experiments. As such, BECON indicators are suited to make the connection between epidemiology and behavior, and thereby allow behavioral scientists to leverage their knowledge and skills in developing optimal interventions to control the pandemic.

The structure of this paper is as follows. First, we will outline the theoretical basis of our approach. Second, we discuss the methodological strategy behind BECON indicators in more detail. Third, we present a simulation study that serves as a proof of concept. Finally, we discuss future extensions of our work.

2 Behavioral interventions and the contact network

To understand the relation between behavior and epidemiology, it is important to introduce an essential mediator in this relation: the contact network. A contact network encodes which people have been sufficiently close to each other to transmit the virus (Newman, 2018; Pastor-Satorras, Castellano, Van Mieghem, & Vespignani, 2015). In contact networks, individuals (or groups of individuals) are represented as nodes, similar to the representation used in well-known social networks. Two nodes are connected by a link if the corresponding individuals

have been in sufficiently close prolonged contact for virus transmission to occur, and disconnected otherwise. Exactly what "sufficiently close" means depends on the virus in question. For Ebola, which is only transmissible through bodily fluids (Drazen et al., 2014), a link in the contact network would mean that the corresponding individuals were in direct physical contact. For SARS-CoV-2, a link could be present when two individuals have been within a distance of 1 or 2 meters of each other for some time, given its airborne transmission (CDC, 2020).

Virus spread on a contact network can be conceptualized as a process in which nodes infect each other via the links in the contact network (Pastor-Satorras et al., 2015). Usually, a closed population is divided into epidemiological "compartments": each individual of the population can be only in one compartment at a time and the compartments describe stages of the disease. Typical examples of compartments include S (susceptible), E (exposed), I (infectious), R (removed, i.e., either cured or deceased) (Keeling & Rohani, 2011). Mathematically, virus spread is a probabilistic process that operates on the contact network topology (Grimmett, 2018; Van Mieghem, 2014) and that specifies infection and curing events, i.e., how long a person is infectious and when the person is cured or deceased. The time distribution of these events, and the local rules at the host (i.e., what happens if a person is in state S, E, I, or R), depend on specifics of the virus in question; for instance, for COVID-19, the consensus during the SARS-Cov-2 variants operative in 2020 held that people were infectious for an average of about 6-7 days (Backer, Klinkenberg, & Wallinga, 2020).

Once the structure of the contact network and the compartment model is specified, the probability or average fraction of individuals in each compartment can be computed per unit time (Sahneh, Scoglio, & Van Mieghem, 2013). If the contact graph does not change too much over time, other global properties of the virus spread can be determined. An important property is the epidemic threshold, which is related to the basic reproduction number R (Pastor-Satorras et al., 2015), and describes the conditions under which outbreaks can occur. If the contact network changes over time, then we enter a complicated situation in which computer simulations are necessary to study virus spread. Another approach is to map the contact graph into a certain class, similar as the classes of Erdős–Rényi or Barabasi-Albert random graphs, and properties of such a contact graph class can be deduced, in principle, analogously (Newman, 2018). Thus, we assume that the time-dependent network has similar properties as the properties of the class and thus abstract the temporal changes in time. Finally, novel approaches based on the analysis of time series data can be used to include the dynamic changes of the network in the analysis (Dekker et al., 2021). In the current paper, we focus on the simplest case, namely one in which the contact network is stable over time so that it can be characterized by a single network structure.

Behavior, contact networks, and compartmental epidemiological models are strongly related: behavior controls the structure of the contact network, the contact network directs the spread of the virus, and the spread of the virus determines population statistics of the epidemiological compartments. Figure 1

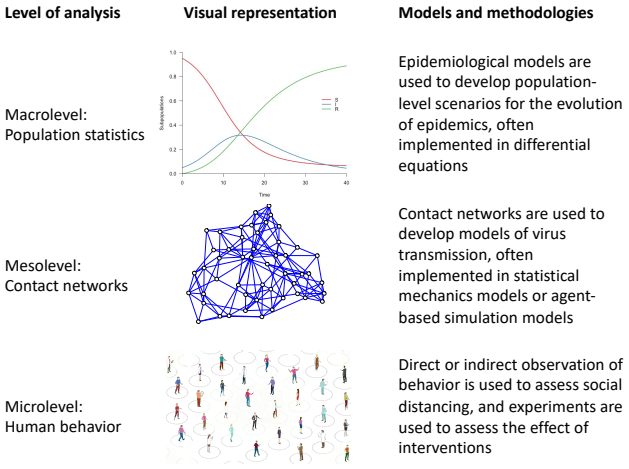


Figure 1. Different levels of analysis that are used in the scientific study of infectious diseases like COVID-19. At the microlevel, human behavior determines physical distances between individuals that are crucial to virus transmission. The resulting pattern of physical distances can be aggregated into a contact network at the mesolevel, in which nodes represent individuals and edges represent physical contacts that make virus transmission possible. These transmission processes determine how many people get infected and at what rate; at the macrolevel, mathematical models based on differential equations are used to model these quantities.

represents this hierarchy of levels visually. This has ramifications for how we should think about behavioral interventions: interventions (e.g., instructing people to practice physical distancing) lead to behavior change (e.g., people will keep more distance), which causes contact networks to change (e.g., the number of links in the network may decrease). These changes cascade into population level statistics (e.g., the value of R will go down) that eventually determine policy success (e.g., number of hospital intakes will stay within limits defined in policy considerations). In accordance, a central idea underlying our framework is that, in order to connect interventions to epidemiological models, they should be represented as operations that transform the contact network (Pastor-Satorras et al., 2015); the present paper applies this idea to behavioral interventions.

This analysis opens up an important methodological possibility: if behavioral interventions operate through changes in the contact network, then measures of that contact network could in principle be used to assess the effect of such interventions. Such an approach would address each of the problems highlighted in the introduction. First, assessment of the contact network can be executed instantaneously, addressing the lagging indicator problem. Second, because the contact network depends only on whether individuals are sufficiently close to each other, and not on whether they actually infect each other, we can potentially pick

up changes in the contact network in the absence of actual virus transmission (of course, applications of insights gained would still require knowledge of how active transmission works). Third, as a consequence of these two properties, assessments of the contact network can be implemented in experimental designs without raising ethical concerns of exposing individuals to virus spread. As a result, the study of interventions would no longer be limited to assessments of policy effects at the societal level (Chu et al., 2020; Flaxman et al., 2020) but could also be used to study manipulations at a much smaller microlevels (e.g., in specific locations like public buildings, restaurants, or concerts). Implementation at the microlevel in turn facilitates the type of controlled experimental research that characterizes psychology (e.g., the implementation of interventions in factorial designs). In the next paragraph, we show how contact networks may be assessed to construct indicators that allow for such approaches.

3 BECON indicator methodology: Strategy and rationale

The idea behind BECON indicators is to assess (functions of) the contact network on the basis of behavioral observations. In the current paper, we will focus on assessments of the contact network using wearables or contact tracing apps that are designed to register whether a person has been within a certain distance (e.g., 1.5 meters) of another person. This methodology has the advantage that it measures a proxy to actual behavior (rather than, e.g., relying on self-reports) and that it does not require a controlled environment, so that it can in principle be used in daily life, enhancing ecological validity.

A schematic of the proposed methodology is represented visually in Figure 2. The interactions between people in the baseline situation (i.e., the situation without the behavioral intervention of interest being implemented) give rise to the true baseline contact network (left bottom).

The true contact network is most likely not directly observable. First, as noted, the presence of a link depends on a theory of virus transmission, which is approximate. For example, in the case of COVID-19, the presence of aerosol transmission or infection via surfaces can create links between people who are at a greater distance than 1.5 meters, or between people who were present at the same place at distinct time points (e.g., because the aerosols remain present in bathrooms after the infectious person has left). Second, if the network integrates contacts over time (e.g., by taking the union of all contact networks at each time point, which registers during a certain time interval who has ever been in contact with who, but not when), the representation will contain false positive connections; for instance, when A and B were in contact, and subsequently B and C were in contact, then the patterns of links suggests that both $A \rightarrow B \rightarrow C$ and $C \rightarrow B \rightarrow A$ are possible infection routes, while only the former route is possible (see also Dekker et al. (2021)). Third, various kinds of measurement errors can yield false positives and false negatives. In a situation in which tracking devices are used, examples of mechanisms that can lead to measurement errors may include hardware failures, signal failures, and failures in data processing.

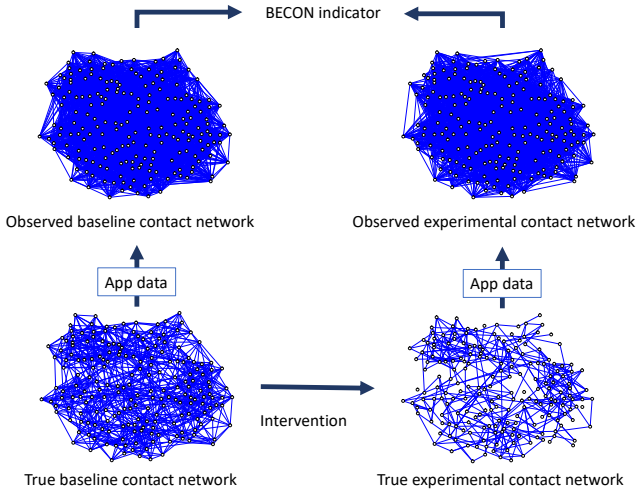


Figure 2. Schematic plan of the proposed methodology. Prior to intervention, the contact network (lower left panel) generates app data that are used to partially reconstruct the network structure (left top panel). Under the intervention, the density of the contact network decreases (lower right panel). This network is also reconstructed on the basis of app data (upper panel). The difference between the reconstructed networks is used to construct a BECON indicator to assess the effect of the intervention.

Thus, the true baseline contact network is generally not observable and difficult to assess adequately. In the present paper we therefore represent this network as a latent structure. To assess this latent structure, we can use observations, as for instance obtained through smartphone apps, video footage analysis, or wearable sensors. For instance, a simple starting point may be to have a group of people use wearable devices that track their location. Measured location data may then be used to decide whether two people have been in contact. Thus, the structure of the data required to construct a contact network is of the form $[AB, BC, \ldots, YZ]$ which encodes that persons A and B, B and C, ..., and Y and Z have been within 1.5 meters of each other for the amount of time specified in the definition of a contact. This is called an edge list in network science, which can be transformed into an adjacency matrix. From this matrix, many important metrics in network analysis can be computed (Newman, 2018). We denote the network encoded in the empirically derived adjacency matrix the observed baseline contact network (top left in Figure 2).

Next, we implement a behavioral intervention. For example, we might instruct people to keep their distance, put up signposts, install an alarm on their phones that sounds when they get too close, or use a variety of nudges that promote physical distancing. If effective, the intervention changes people's behavior, and as such induces changes in the contact network. We denote the resulting network as the true experimental contact network. Like the baseline

contact network, the true experimental contact network is not directly observable, but can be assessed indirectly through location measurements. Such data may again be obtained by letting people use wearables in the experimental situation, after which the results can be used to arrive at an observed experimental contact network.

Recall that a behavioral intervention effect is, in essence, a transformation of the contact network. Hence, if we could directly assess the baseline and experimental contact networks, we could precisely determine the effect of the intervention and, given a dynamical regime, we could also assess the degree to which the intervention should be expected to mitigate virus spread. Unfortunately, however, we cannot directly compare the baseline and experimental networks, as these are only indirectly observable. However, we can directly compare the observed networks created through measurements. For example, one could compute the number of links in the observed experimental contact network, and compare that quantity to the number of links in the observed baseline network. This way, one could assess whether, on average, people keep more distance in the experimental condition. From a bird's eye perspective, observational studies have shown a substantial decline in average mobility after lockdowns have been enforced (e.g., Jeffrey et al., 2020). However, one could also utilize a variety of more advanced network metrics which can provide a detailed picture of the changes that the intervention has produced. For example, one could compare the networks for their density, diameter, average shortest path lengths, etc. to assess the effect of specifically targeted interventions (e.g., interventions that target specific individuals central in the contact network, like doctors and teachers).

The function that is chosen to assess the difference between networks defines a Behavioral Contact Network indicator; a BECON. In this paper, we develop three BECONS: The Density BECON, the Spectral BECON, and the ASPL BECON. To facilitate interpretation, each of the BECON indicators is constructed in such a way that a higher value on the indicator implies that the experimental network has changed the network in a direction that would in typical circumstances be expected to limit the potential for a virus to spread. The indicators studied here are are defined as follows.

The Density BECON uses the relative change in network density. Network density is defined as the ratio of the number of links to the total number of possible links in the network. This measure is epidemiologically relevant, because denser networks indicate that more people have been in close proximity to each other. The Density BECON is constructed by dividing the density of the observed baseline contact network by that of the observed experimental contact network, where higher values indicate larger experimental effects. In essence, this measure simply tracks the extent to which the number of contacts reduces in the experimental condition. This measure would be most relevant in microlevel applications, where people for instance use wearables during a public event, because in this case only the direct contacts are relevant. This is because for COVID-19, a person will take several days from infection to being infectious; hence, indirect connections that would lead to transfer from one person to another via a third

person ($A \to B \to C$) are not possible on such short time-scales. This metric is particularly suitable in microlevel applications when the period from infection to being infectious takes multiple days, as was the case with COVID-19. In this situation, indirect connections that would leads to transfer from one person to another via a third person ($A \to B \to C$) are not possible and, hence, the (dynamical) contacts can be concatenated into a single contact network.

The Spectral BECON is based on the spectral radius. The spectral radius is the largest eigenvalue of the adjacency matrix. The spectral radius is epidemiologically important, because the inverse of the spectral radius is equal to the so-called mean-field epidemic threshold, which is in turn a lower bound to the real epidemic threshold (Van Mieghem & Van de Bovenkamp, 2013; Van Mieghem & van de Bovenkamp, 2015). The epidemic threshold plays a central role in networks and copes with structural heterogeneity, because a viral strength above the epidemic threshold will endemically infect a non-zero fraction of the nodes in the network. In the limiting case of a complete graph on N nodes, where all nodes are connected to each other, the epidemic threshold is approximately equal to $1/N$ and the strength of the virus divided by the epidemic threshold is approximately equal to the reproduction number, whose critical value is equal to one (for a reproduction number larger than one, the model predicts the virus to be endemic, while for values below one it predicts that the virus will eventually disappear). Moreover, one may control and tune the contact network so that its spectral radius is minimized and the vulnerability for infections (virus spread) is maximized (Van Mieghem et al., 2011). The Spectral BECON is constructed by dividing the largest eigenvalue of the adjacency matrix of the observed baseline contact network by that of the observed experimental contact network, such that a larger value indicates a larger intervention effect. The Spectral BECON would not be relevant in microlevel research (e.g., tracking people in a location over several hours), but rather would apply to measures taken over days, as could be gained using contact tracing apps. A specific example would be contact tracing within the workspace, where the same group of people met each other over longer periods of time. Different set-ups and interventions could be tried out (e.g., different 'bubbles' of employees, different organisation of common spaces, walking routes), and their effectiveness could be assessed using the Spectral BECON. The Spectral BECON is thus useful to assess intervention effects that involve changes in the network structure that not only affect the number of links per node, but work on the architecture of the network as a whole.

The ASPL BECON uses the average shortest path length (ASPL) between pairs of nodes in the network. The shortest path length (SPL) between two nodes equals the minimum number of edges that one has to traverse to travel from one node to the other; the ASPL is the average value of all SPLs between all pairs of network nodes. The ASPL is relevant to virus transmission, because the shorter the paths that connect nodes in a contact network are, the easier the virus can spread from one person to randomly chosen other person. The ASPL BECON is constructed by dividing the ASPL of the observed experimental contact network by that of the observed baseline contact network. Like the Spectral BECON,

the ASPL BECON could be applied in research where contacts are traced over a period of days, in which indirect connections ($A \rightarrow B \rightarrow C$) are relevant. For calculation of the ASPL BECON we ignore any infinite shortest path lengths that arise from disconnected graphs (i.e., in the case when some nodes or groups of nodes are unconnected). Thus if there was more than one single connected component (which happened in less than 3% of the cases), paths between nodes in different connected components did not exist. We hence ignored these when computing the ASPL BECON.

The fact that each is expressed as a ratio allows one to interpret the BECON values directly: for instance, if the Density BECON equals 2, this means that the density of the observed baseline contact network is twice as large as that of the observed experimental contact network. Other things being equal, a higher BECON value would indicate that the intervention would likely be more successful in mitigating virus spread (naturally, this should be considered in the light of a theory about the virus transmission process). In research at the microlevel, where people are traced over periods of hours, the Density BECON would be most relevant.

An important methodological question is whether we can use BECON indicators to assess the effect of interventions in realistic circumstances, where our assessments of the contact network will be distorted in various ways. Thus, the question that arises is whether the setup sketched in Figure 2 can be used to assess the effect sizes of experimental manipulations in realistic conditions. For example, can one order the effects of a set of behavioral interventions in terms of effect size? How does the methodology fare in the presence of realistic amounts of measurement error? To assess whether this is indeed possible, we now turn to a simulation study.

4 Simulation study

The simulation study is designed to evaluate whether the BECON methodology is indeed able to pick up effects of interventions if these are present. To show this, we vary the size of intervention effects on the true contact network, and subsequently assess the corresponding BECON values in the observed contact network which is subjected to various levels of noise. If the method is reliable, we expect the BECON values to be higher if the effects are stronger. We evaluate this by computing the correlation between the size of the simulated intervention effect and the observed BECON values: the higher the correlation is, the more reliable the BECON is as an indicator of intervention effects.

In the simulation, we vary the number of nodes in the network $n \in [100, 200, 500, 1000]$, specifying the architecture of the contact network as a small world structure (Watts & Strogatz, 1998) and generate it using the R package *igraph* (Csardi & Nepusz, 2006), with a neighborhood size of $K = 5$ and a rewiring probability of $p = 0.10$. We use this specification because it leads to a network structure with a high degree of clustering yet small average shortest path lengths, which is qualitatively similar to the structure of some social networks found in

empirical research, where links may e.g., encode whether individuals work at the same place or visit the same sports club. This network serves as our baseline contact network (Figure 2, left bottom).

Next, we simulate a measurement process (Figure 2, left arrow). The process generates imperfect measures of the baseline contact network. We simulated a measurement function with a false negative rate of $fn = [0.10, 0.20, 0.30]$ (e.g., $fn = 0.30$ implies that only 70% of the network links were successfully picked up) and false positive rate $fp = [0.10, 0.20, 0.30]$ (e.g., $fp = 0.10$ implies that 10% of the links that are absent in the true contact network will be present in the observed network). We emphasize that, in the present paper, our interest is not to retrieve the actual network structure by correcting for these distortions or to recover the dynamical processes that it supports. Instead, our primary interest here lies in the pragmatic goal of assessing the effect of interventions, so as to develop an indicator that can serve to bridge the gap between epidemiology and behavioral science.

Subsequently, we assess the observed baseline contact network using the obtained data (Figure 2, top left). This network can be quite severely distorted as a result of the probabilistic nature of the measurement function. For instance, as is visible in the figure, the observed network is much denser than the true contact network, even though the false positive rate is much lower than the false negative rate. This is because the true contact network is relatively sparse, which means it contains more absent than present links: the small world networks we generated for $n = 100$ individuals have $n \times (n-1)/2 = 4,950$ possible links, of which on average only $K \times n = 500$ are actually present. Hence, a false positive rate of 10% generates about $0.10 \times (4,950 - 500) = 445$ false positive links, while the false negative rate of 30% means that on average only 350 of the 500 actual links are successfully identified. In other words, the observed baseline contact network contains about $445 + 350 = 795$ links, of which only 350 are true positives. Because links feature such a low base rate, the probability that two nodes that are connected in the observed network are actually connected in the true contact network is only 350/795, i.e., about 0.44. This effect is stronger in larger networks, because in larger networks there are more opportunities for false positives; for example, in a network of 1000 individuals, the probability that an observed link is actually present in the true network is only 0.07. Thus, the actual assessment of the network in itself may be largely unsuccessful, especially in larger networks. We think that similar results would be expected in actual empirical work if the studied contact networks are sparse. However, as we will see, the lack of success in assessing the contact network itself does not preclude the possibility of assessing intervention effects by comparing the observed networks.

We now implement an intervention on the network (Figure 2, bottom arrow). A typical intervention would be intended to, e.g., improve physical distancing, and hence should lead to a lower connectivity in the experimental contact network (Figure 2, bottom right). We model such an intervention by deleting links uniformly at random (a process known as bond percolation in the network literature; Newman, 2018). The proportion of links deleted from the baseline con-

tact network then defines the effect size of the intervention. We vary this effect size in steps of 10%, from an ineffective intervention, which deletes no links, to the strongest intervention, which deletes 90% of the links. We implement this intervention in two ways: the deterministic intervention removes the specified proportion of links exactly, by randomly deleting links until this proportion is reached, while the probabilistic intervention removes each link with a probability equal to the specified proportion. Thus, in an intervention setting with effect size of, say, 0.40, the deterministic intervention results in a network where exactly 40% of the links are deleted, while the probabilistic intervention removes each link with a probability of 0.40, so that the expected percentage of removed links equals 40% while each realization may be different. The probabilistic intervention thus implements a situation in which the interventions are represented as a random effect that differs across experimental settings, leading to more uncertainty. Importantly, these interventions are but two of the many alternative and possibly directed interventions that one may study (Trajanovski, Martín-Hernández, Winterbach, & Van Mieghem, 2013).

Finally, we implement the same measurement function as before on the experimental contact network. This way, we arrive at the observed experimental network (Figure 2, top right). As was the case for the baseline contact network, this assessment is dominated by false positives, which leads to a significant overestimation of the network density. In addition, the fact that the experimental contact network is sparser than the baseline contact network implies that the probability of a link being present in the experimental network, given that it is present in the observed experimental network, has diminished even more.

Table 1. Simulation results across conditions for a small world graph. The table reports correlations (means and sd) between BECONs and intervention effect sizes for deterministic and probabilistic interventions across simulation runs for multiple network sizes. The results are averaged over the false negative rates.

n	FP	Deterministic						Probabilistic					
		Density		Spectral		ASPL		Density		Spectral		ASPL	
		r	SD	r	SD	r	SD	r	SD	r	SD	r	SD
100	0.1	1	0	1	0	1	0	0.97	0.03	0.96	0.04	0.97	0.03
100	0.2	1	0	1	0	1	0	0.92	0.05	0.92	0.06	0.92	0.05
100	0.3	1	0	1	0	1	0	0.86	0.10	0.86	0.10	0.85	0.11
200	0.1	1	0	1	0	1	0	0.97	0.03	0.96	0.03	0.97	0.02
200	0.2	1	0	1	0	1	0	0.92	0.06	0.91	0.06	0.91	0.06
200	0.3	1	0	1	0	1	0	0.86	0.10	0.85	0.10	0.85	0.11
500	0.1	1	0	1	0	1	0	0.96	0.03	0.96	0.03	0.97	0.02
500	0.2	1	0	1	0	1	0	0.92	0.07	0.92	0.06	0.91	0.07
500	0.3	1	0	1	0	1	0	0.85	0.11	0.84	0.11	0.84	0.11
1000	0.1	1	0	1	0	1	0	0.96	0.03	0.96	0.03	0.96	0.03
1000	0.2	1	0	1	0	1	0	0.91	0.06	0.91	0.06	0.91	0.06
1000	0.3	1	0	1	0	1	0	0.85	0.10	0.85	0.10	0.85	0.10

As a measure of the accuracy of the BECONs in ordering the intervention effect sizes, we compute the average correlation between the estimated BECONs and the actual intervention effect across simulation runs. A correlation of unity then means that the BECONs order the interventions perfectly, while a correlation of 0 means that the BECONs do no better than chance.

Results of the simulations are given in Table 1. As can be seen from the table, despite the fact that the networks used to compute the BECONs were poor representations of the actual contact networks, the difference between the observed networks tracks the intervention effect size reliably. To illustrate how these effects arise, Figure 3 gives a detailed representation of results for one specific BECON in the simulation design, i.e., the ASPL BECON performance with $fp = 0.20$ and $fn = 0.20$. Detailed representations of simulation results for all BECONs across all conditions are given in the Appendix. As shown in detail in the Appendix, the results in Figure 3 are representative of the behavior of all three BECONs, which uniformly varied as a monotonic function of effect size, in the sense that the probability distributions of these statistics stochastically order the interventions. The separation of effect sizes is in fact perfect for all deterministic intervention simulations. In the probabilistic intervention simulation, results are somewhat more attenuated, but the correlation between true and estimated intervention effects does not drop below 0.80. Thus, even with sizeable false positive and false negative rates, the methodology still works extremely well.

As can be seen in Figure 3 and the extended results in the Appendix, BECONs are monotonically related to intervention effect sizes, such that stronger effects result in higher BECONs. Yet, for larger networks, the increase in BECON attenuates (i.e., the slope becomes smaller). This can be explained by the small world structure of the networks, which makes larger networks relatively more sparse compared with smaller networks. As a result, a fixed false positive rate in the measurement process (e.g., 20% false positives) has a more pronounced effect in the larger networks. This is because the number of present links grows linearly with the number of nodes, while the number of possible links grows quadratically, and therefore larger networks feature a smaller percentage of present links. For example, a neighborhood size of 5 results in the presence of about 1% of the possible links in a network of 1000 people ($5,000$ out of $449,500$), but the presence of about 10% of the possible links in a network of 100 people (500 out of $9,900$). As this measurement process applies to both the observed baseline network and to the observed experimental network, the false positive rate will make larger networks relatively more similar to each other than smaller networks. Consequently, it is more difficult to detect an effect in larger networks, which is reflected in the BECONs. However, as is evident from Table 1, BECON performance is robust against this effect across the conditions simulated.

Finally, we find essentially the same results for a scale-free network, that is, a network whose degree distribution follows a power law, and for an Erdős–Rényi random graph, which has a binomial degree distribution. Results for these network structures as well as code to reproduce them are available at `https://`

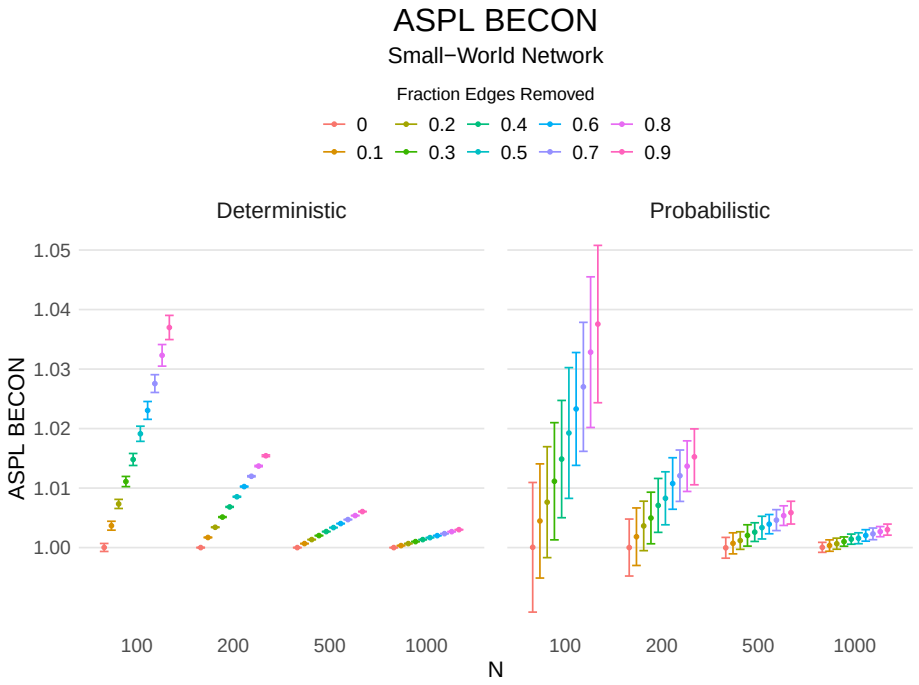


Figure 3. ASPL BECON indicator values as a function of intervention effect sizes. Plots display the range that contains 95% of the observed BECON values in the relevant condition. The true network structure is a small-world network. This figure displays results for a setting with false positive and false negative rate of 0.20

`gitlab.com/science-versus-corona/becon`. Performance is broadly consistent across network structures. The reason for this may lie in the noisy measurement process: the false positive rate results in a substantial amount of new links, which obscure the original structure and thereby may counter effects of network structure.

5 Discussion

In this paper, we have introduced a behavioral data science methodology to study interventions designed to counter virus spread, and have shown in simulations that this methodology is both feasible and effective. BECON indicators, constructed to pick up changes in the contact network that are induced by interventions, were shown to track intervention effects reliably under realistic measurement conditions that are characterized by substantial levels of noise. Thus, BECON indicators offer a promising methodological approach to studying interventions designed to mitigate virus spread in COVID-19 and other infectious diseases.

Because BECON indicators are instantaneous and can be constructed in the absence of actual virus spread, they address the main problems that plague widely used indicators, such as hospital admissions and death rates. In contrast to current indicators, BECON indicators do not suffer from lags and can be used in experimental designs; in addition, if used in the absence of actual virus transmission, the use of BECON indicators need not put participants in control conditions at risk. Therefore, BECON indicators have the important advantage that they can be used when viruses are not present, i.e., they allow us to maximize our defenses and to prepare for new outbreaks. Of course, the interpretation of the results does depend on details of the virus transmission process, and an important question is how to relate changes in the observed contact networks to this process.

As indicated, the study of which BECON is best suited for a given research question depends on a combination of factors including characteristics of the virus, properties of the research design, and the goals of the research program. One important issue, highlighted throughout this manuscript, involves the time scale at which the research program runs in relation to the time scale at which a virus spreads. In cases where people are observed for a shorter time than that needed for the virus to incubate and become infectious, the Density BECON is always best, because the virus cannot travel more than a single step in the network. In cases where people are followed for a period that exceeds this period, the Spectral and ASPL BECONs become feasible alternatives. If one wants to study effects of generic interventions (e.g., lockdowns or school closures) on virus spreading potential, then the Spectral BECON is indicated due its close relation to the epidemic threshold (Van Mieghem & Van de Bovenkamp, 2013; Van Mieghem & van de Bovenkamp, 2015). Finally, in cases where one has interventions that are specifically targeted at the path lengths in the network (e.g., interventions targeted to limit interactions based on a previous interaction history), one can use the ASPL BECON. Thus, generally, we would recommend the Density BECON in all situations where the research design observes behavior at a duration below that needed for the virus to spread, and the Spectral and ASPL BECONs in research designs that observe behavior for longer durations; the choice between Spectral and ASPL BECONs then depends on specifics of the research question.

The Density BECON offers a simple metric to be used in small scale experiments, where one for instance wants to test the effectiveness of office designs in a company, or where one desires to assess the relative effect of different nudges. We performed a study in which we applied the BECON methodology in practice. During an art fair, we implemented different nudges and evaluated its effect on the contact network. This way we could for example show that walking directions positively impacted physical distancing and reduced the number of contacts, demonstrating the effectiveness of the proposed methodology in practice (Blanken et al., 2021; Tanis et al., 2021). Simulations indicate that even for small networks the indicators are reliable. However, BECON indicators are potentially also applicable to large scale research. For example, using contact tracing apps,

it should be possible to assess contact networks at the scale of neighborhoods, cities, or even countries. Thus, BECON indicators could be implemented in a dashboard used by policy makers to assess the degree to which current policies are on track. Because they are much faster than traditional indicators, they may also be highly useful in contributing to alarm systems that indicate that policy action is required.

While our approach aims to change the network structure by deleting links, a closely related notion is the removal of nodes (known as site percolation; Newman, 2018). This is especially important in vaccination campaigns, especially if vaccinating an individual leads to a situation where the vaccinated individual cannot receive nor spread the disease (Y. Liu et al., 2021).[3] Strictly speaking, vaccinating an individual does not change the network structure. For the purposes of disease spreading, however, one may reformulate it as such: in cases where vaccinating an individual ensure that the individual will no longer be able to spread the disease, this implies that all links going into and out of the node will be removed. Although it focuses on node rather than link removal, research done on animal populations is related to the approach we propose here. Carne, Semple, Morrogh-Bernard, Zuberbuehler, and Lehmann (2013), for example, use simulations to study how targeted vaccination (e.g., vaccinate the most central nodes) outperform random vaccination in a network derived from observations of orangutans and chimpanzee populations, respectively. They used cluster size and shortest paths of the network as measures to assess the effect of interventions (see also Albert, Jeong, & Barabási, 2000). Relating the network structure to the final outbreak size, Rushmore et al. (2014) find in simulations that vaccinating the most connected chimpanzees can reduce the outbreak size considerably. While these articles focus on node removal, we expect that our approach can learn from such approaches; future research may explore this line more fully.

Several limitations to the present work should be noted. For example, as we have emphasized throughout this paper, observed networks will ordinarily be poor representations of the contact networks of interest unless contact network assessments are augmented by more advanced measurement methods. Therefore, one should be very careful in using observed contact networks as proxies for the underlying contact networks. For example, the fact that observed networks are likely to be much more dense than the underlying contact networks suggests that it would not be a good idea to use these observed networks naively in, e.g., simulations of virus transmission. However, our primary interest in this paper was not in the reconstruction of contact networks per se, but in the comparison of contact networks across experimental conditions. Standard experimental wisdom holds that errors in observations need not form an insurmountable problem as long as the structure and size of the induced distortions is comparable across experimental conditions; in this case, systematic errors will be invariant across conditions, and random errors will average out as the number of observations

[3] Here it is important to note that for COVID-19, vaccination does not preclude an individual from receiving or spreading the virus, although transmission rates among vaccinated individuals were reduced. (Eyre et al., 2022)

grows. This indeed is shown to be the case in our simulations. It should be noted that more advanced reconstruction methods may lead to biases if they are differentially effective across conditions, so a better reconstruction need not imply a better signal of intervention effectiveness.

Our approach invites improvements at several points. First, we study a simple measurement process. In the real world, it is likely that the observed network is not biased in a manner as we study here (i.e., by adding false positives and false negatives), but that more complicated biases occur as would, for example, arise when the structure of missing data depends on the network structure itself. Such biases can be addressed by adding an intermediary network reconstruction step. In particular, before computing the BECONs, one can use network reconstruction algorithms to first arrive at a better representation of the true network (e.g., Clauset, Moore, & Newman, 2008; Ghasemian, Hosseinmardi, Galstyan, Airoldi, & Clauset, 2020; Goyal & Ferrara, 2018; Guimerà & Sales-Pardo, 2009). Similarly, the use of additional sources of information deliverable through mobile phones (e.g., geographical location data, WIFI data, ultra wide-band technology) could enhance the precision of the signal (Trofimenko, Mukhina, & Visheratin, 2016). In addition, if the amount of bias in the measurement function is not equal across experimental conditions, the presented methodology would likely lead to incorrect conclusions. Because of the sparsity of the contact networks, even differences in random noise could potentially lead to bias in the effect sizes under certain conditions. For example, if the experimental intervention increases the percentage of false negatives, it can seem effective while it is not. Statistical corrections could be developed on the basis of latent variable models, which are able to accommodate violations of measurement invariance to some extent (Meredith, 1993; Van De Schoot et al., 2013).

Another open question is whether the validity of BECONs is symmetric; can we pick up interventions that make the network more densely connected as easily as interventions that prune it? This is an important question in the process of monitoring lifting regulations, which is expected to create increasingly connected contact networks. BECON methodology could be used to assess the relative risk of different lifting interventions experimentally, and as such may inform exit strategies.

Given that the value of the BECON methodology especially lies in its ability to tap into actual (distancing) behavior, we hope the present paper contributes to experimental research into the effectiveness of behavioral interventions in this domain. However, the interventions we have studied in this paper are very simple, as they delete links at random. One may interpret such an intervention as reducing contacts at random. It would be interesting to investigate what happens if an intervention does not randomly delete links, but affects the structure of the network in a different way (as for example in Trajanovski et al., 2013). For example, one could examine what happens if interventions selectively take out shortest paths, but keep clusters (e.g., groups of friends) intact. This would probably change not only the density, but also the structure of the contact network after intervention; for example, selectively targeting nodes with high centrality

may lead the network to lose its small world character. These interventions would have considerable effect on epidemic spread. One such investigation was recently provided by Block et al. (2020). The authors study three types of interventions — limiting social interactions to a few individuals, seeking similarity across contacts, and strengthening communities — that change the contact network. They subsequently simulate virus spread and find that all three interventions substantially flatten the infection curve compared to no intervention, as well as to an intervention that makes actors randomly reduce their contacts (i.e., removes links at random). As suggested above, this shows that more thoughtful interventions can have a more drastic effect on virus transmission. Since different interventions change the contact network in different ways, it is important to choose the right BECON. In addition, such thought experiments suggest that the question under which conditions which BECONs can adequately track virus transmission is open for future research; one could imagine, for instance, implementing different interventions and running epidemiological virus spread models on the resulting networks to understand how different interventions change the epidemiological course of the virus. Finally, the current setup ignores dynamical information about interventions (e.g., the duration of effects), and extending measurements and models in this direction could augment the signal considerably (Dekker et al., 2021). Such information could then be used to assess these interventions in a more precise fashion.

Not all interventions are amenable to the BECON approach. For instance, certain interventions may be highly effective in controlling the virus spread without implementing large changes in the network. A salient example arises when interventions are directed at interrupting processes that run on specific parts of the contact network, rather than at changing the network structure globally. Such interruptions are, for instance, the goal of contact tracing (Cencetti et al., 2020; Kojaku, Hébert-Dufresne, Mones, Lehmann, & Ahn, 2021; Kretzschmar et al., 2020). Interventions based on contact tracing provide highly local and surgical interventions on the network (namely, by isolating potentially infected cases, such procedures delete the corresponding links from the contact network). It is unlikely that global measures like BECONs would be able to pick up such subtle effects, especially if cases are rare so that relatively few links are deleted. Also, it would be important to study whether contact tracing interventions invariably lead to structures that diminish the potential for virus transmission; one can imagine situations where alarm systems based on contact tracing may instigate behavior that increases virus transmission. This would especially be relevant if alarm systems are unreliable or are activated too late, which underscores the importance of accurate prediction.

In the future, network theory may assist behavioral scientists in developing novel interventions. For example, if advanced reconstruction of the contact network to a high level of precision becomes feasible, interventions could be shaped by the analysis of the baseline network itself. In such an approach, one could first analyze the baseline contact network, and then explicitly design interventions to target particular aspects of the contact network to induce maximal

change, analogous to the use of targeted vaccinations in epidemiology (Q. Liu, Zhou, & Van Mieghem, 2019; Pastor-Satorras & Vespignani, 2002). Similarly, interventions could be explicitly targeted to decrease the spectral radius of the contact network (Van Mieghem et al., 2011), because this controls the potential for outbreaks (Van Mieghem & Van de Bovenkamp, 2013; Van Mieghem & van de Bovenkamp, 2015). In accordance, targeted evaluations of changes in the contact network after intervention could be used to assess whether the intended changes have indeed been accomplished. This approach potentially defines an extensive research program, in which behavioral data scientists, epidemiologists, psychologists, computer scientists, and statisticians could profitably work together to construct, implement, and monitor optimal interventions.

References

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, *406*(6794), 378–382. doi: https://doi.org/10.1038/35019019

Backer, J. A., Klinkenberg, D., & Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019- nCoV) infections among travellers from Wuhan, China, 20 28 January 2020. *Eurosurveillance*, *25*(5), 1–6. Retrieved from `http://dx.doi.org/10.2807/1560-7917.ES.2020.25.5.2000062` doi: https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062

Blanken, T. F., Tanis, C. C., Nauta, F. H., Dablander, F., Zijlstra, B. J. H., Bouten, R. R. M., ... Borsboom, D. (2021). Promoting physical distancing during COVID-19: a systematic approach to compare behavioral interventions. *Scientific Reports*, *11*(1), 19463. Retrieved from `https://doi.org/10.1038/s41598-021-98964-z` doi: https://doi.org/10.1038/s41598-021-98964-z

Block, P., Hoffman, M., Raabe, I. J., Dowd, J. B., Rahal, C., Kashyap, R., & Mills, M. C. (2020). Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world. *Nature Human Behaviour*, 1–9. doi: https://doi.org/10.1038/s41562-020-0898-6

Carne, C., Semple, S., Morrogh-Bernard, H., Zuberbuehler, K., & Lehmann, J. (2013). Predicting the vulnerability of great apes to disease: the role of superspreaders and their potential vaccination. *PLoS One*, *8*(12), e84642. doi: https://doi.org/10.1371/journal.pone.0084642

CDC. (2020). *How COVID-19 spreads.* Retrieved 2021-07-14, from `https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html` (Accessed on 28-09-2021)

Cencetti, G., Santin, G., Longa, A., Pigani, E., Barrat, A., Cattuto, C., ... Lepri, B. (2020). Using real-world contact networks to quantify the effectiveness of digital contact tracing and isolation strategies for Covid-19 pandemic. *medRxiv*.

Chu, D. K., Akl, E. A., Duda, S., Solo, K., Yaacoub, S., Schünemann, H. J., ... others (2020). 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*, *395*(10242), 1973–1987. doi: https://doi.org/10.1016/j.jvs.2020.07.040

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*(7191), 98–101. doi: https://doi.org/10.1038/nature06830

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5), 1–9.

Dekker, M. M., Blanken, T. F., Dablander, F., Ou, J., Borsboom, D., & Panja, D. (2021). Quantifying agent impacts on contact sequences in social interactions. *arXiv preprint arXiv:2107.01443*. doi: https://doi.org/10.1038/s41598-022-07384-0

Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T., & Jacobsen, K. H. (2019). Complexity of the basic reproduction number (R_0). *Emerging infectious diseases*, *25*(1), 1.

de Vlas, S. J., & Coffeng, L. E. (2021). Achieving herd immunity against covid-19 at the country level by the exit strategy of a phased lift of control. *Scientific Reports*, *11*(1), 1–7. doi: https://doi.org/10.1038/s41598-021-83492-7

Drazen, J., Kanapathipillai, R., Campion, E., Rubin, E., Hammer, S., Morrissey, S., & Baden, L. (2014). Ebola and quarantine. *The New England Journal of Medicine*, *371*(21), 2029–2030. doi: https://doi.org/10.1056/nejme1413139

Eyre, D. W., Taylor, D., Purver, M., Chapman, D., Fowler, T., Pouwels, K. B., ... Peto, T. E. (2022). Effect of covid-19 vaccination on transmission of alpha and delta variants. *New England Journal of Medicine*, *386*(8), 744-756. Retrieved from `https://doi.org/10.1056/NEJMoa2116597` doi: https://doi.org/10.1056/NEJMoa2116597

Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... others (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. *Imperial College London*, *10*(77482), 491–497.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., ... others (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, *584*(7820), 257–261. doi: https://doi.org/10.1038/s41586-020-2405-7

Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airoldi, E. M., & Clauset, A. (2020). Stacking models for nearly optimal link prediction in complex networks. *Proceedings of the National Academy of Sciences*, *117*(38), 23393–23400. doi: https://doi.org/10.1073/pnas.1914950117

Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, *151*, 78–94. doi: https://doi.org/10.1016/j.knosys.2018.03.022

Grimmett, G. (2018). *Probability on graphs: random processes on graphs and lattices* (Vol. 8). Cambridge University Press. doi: https://doi.org/10.1017/9781108528986

Guimerà, R., & Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, *106*(52), 22073–22078. doi: https://doi.org/10.1073/pnas.0908366106

Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., ... others (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*(6227). doi: https://doi.org/10.1126/science.aaa4339

Jeffrey, B., Walters, C. E., Ainslie, K. E., Eales, O., Ciavarella, C., Bhatia, S., ... others (2020). Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with COVID-19 social distancing interventions was high and geographically consistent across the UK. *Wellcome Open Research*, *5*. doi: https://doi.org/10.12688/wellcomeopenres.15997.1

Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press. doi: https://doi.org/10.2307/j.ctvcm4gk0

Kojaku, S., Hébert-Dufresne, L., Mones, E., Lehmann, S., & Ahn, Y.-Y. (2021). The effectiveness of backward contact tracing in networks. *Nature Physics*, *17*(5), 652–658. doi: https://doi.org/10.1038/s41567-021-01187-2

Kretzschmar, M. E., Rozhnova, G., Bootsma, M. C., van Boven, M., van de Wijgert, J. H., & Bonten, M. J. (2020). Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet Public Health*, *5*(8), e452–e459. doi: https://doi.org/10.1016/s2468-2667(20)30157-2

Liu, Q., Zhou, X., & Van Mieghem, P. (2019). Pulse strategy for suppressing spreading on networks. *Europhysics Letters*, *127*(3), 38001. doi: https://doi.org/10.1209/0295-5075/127/38001

Liu, Y., Sanhedrai, H., Dong, G., Shekhtman, L. M., Wang, F., Buldyrev, S. V., & Havlin, S. (2021). Efficient network immunization under limited knowledge. *National Science Review*, *8*(1), nwaa229.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi: https://doi.org/10.1007/bf02294825

Newman, M. (2018). *Networks*. Oxford university press. doi: https://doi.org/10.1093/oso/9780198805090.001.0001

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, *87*(3), 925–979. doi: https://doi.org/10.1103/revmodphys.87.925

Pastor-Satorras, R., & Vespignani, A. (2002). Immunization of complex networks. *Physical review E*, *65*(3), 036104. doi: https://doi.org/10.1103/physreve.65.036104

Rushmore, J., Caillaud, D., Hall, R. J., Stumpf, R. M., Meyers, L. A., & Altizer, S. (2014). Network-based vaccination improves prospects for disease control in wild chimpanzees. *Journal of the Royal Society Interface*, *11*(97),

20140349. doi: https://doi.org/10.1098/rsif.2014.0349

Sahneh, F. D., Scoglio, C., & Van Mieghem, P. (2013). Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking*, *21*(5), 1609–1620. doi: https://doi.org/10.1109/tnet.2013.2239658

Tanis, C. C., Leach, N. M., Geiger, S. J., Nauta, F. H., Dablander, F., Harreveld, F. v., ... Blanken, T. F. (2021). Smart Distance Lab's art fair, experimental data on social distancing during the COVID-19 pandemic. *Scientific Data*, *8*, 179. doi: https://doi.org/https://doi.org/10.1038/s41597-021-00971-2

Trajanovski, S., Martín-Hernández, J., Winterbach, W., & Van Mieghem, P. (2013). Robustness envelopes of networks. *Journal of Complex Networks*, *1*(1), 44–62. doi: https://doi.org/10.1093/comnet/cnt004

Trofimenko, T. B., Mukhina, K. D., & Visheratin, A. A. (2016). Mobile contacts network reconstruction using call domain records data. In *2016 third european network intelligence conference (enic)* (pp. 55–60). doi: https://doi.org/10.1109/enic.2016.016

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*, 770. doi: https://doi.org/10.3389/fpsyg.2013.00770

Van Mieghem, P. (2014). *Performance analysis of complex networks and systems*. Cambridge University Press.

Van Mieghem, P., Stevanović, D., Kuipers, F., Li, C., Van De Bovenkamp, R., Liu, D., & Wang, H. (2011). Decreasing the spectral radius of a graph by link removals. *Physical Review E*, *84*(1), 016101. doi: https://doi.org/10.1103/physreve.84.016101

Van Mieghem, P., & Van de Bovenkamp, R. (2013). Non-Markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks. *Physical review letters*, *110*(10), 108701. doi: https://doi.org/10.1103/physrevlett.110.108701

Van Mieghem, P., & van de Bovenkamp, R. (2015). Accuracy criterion for the mean-field approximation in susceptible-infected-susceptible epidemics on networks. *Physical Review E*, *91*(3), 032812. doi: https://doi.org/10.1103/physreve.91.032812

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440–442. doi: https://doi.org/10.1515/9781400841356.301

Appendix

A Simulation results

Additional tables and figures reporting the simulation results across all conditions. Complete results are available at `https://gitlab.com/science-versus-corona/becon`.

A.1 Small world graph

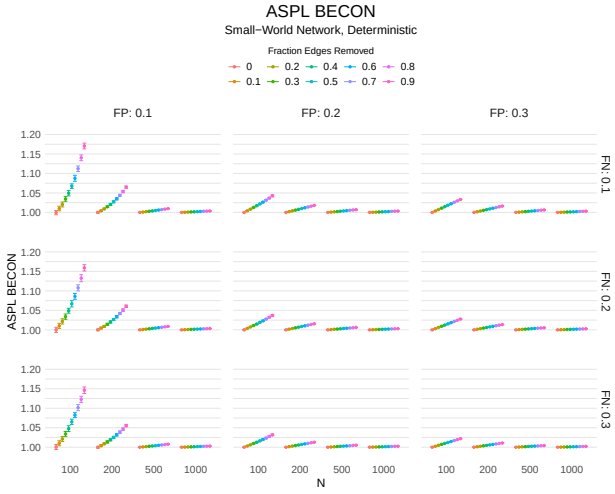


Figure 4. ASPL BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a small-world network. A fixed number of false positives and negatives were added in each run to create the observed network.

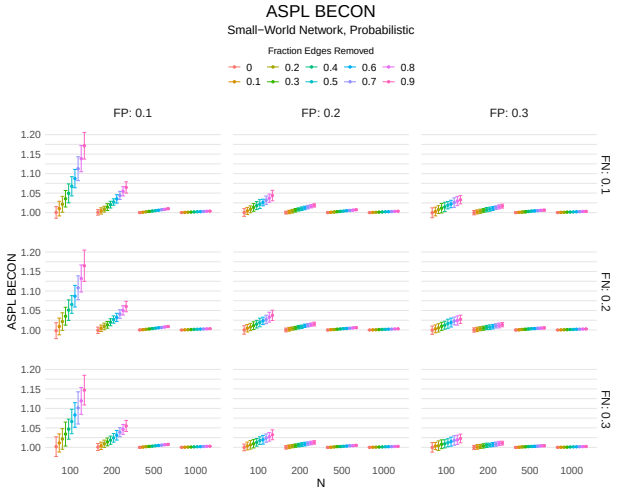


Figure 5. ASPL BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a small-world network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

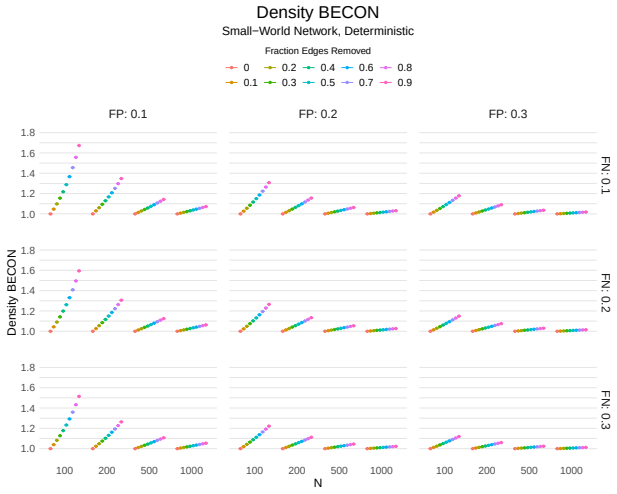


Figure 6. Density BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a small-world network. A fixed number of false positives and negatives were added in each run to create the observed network.

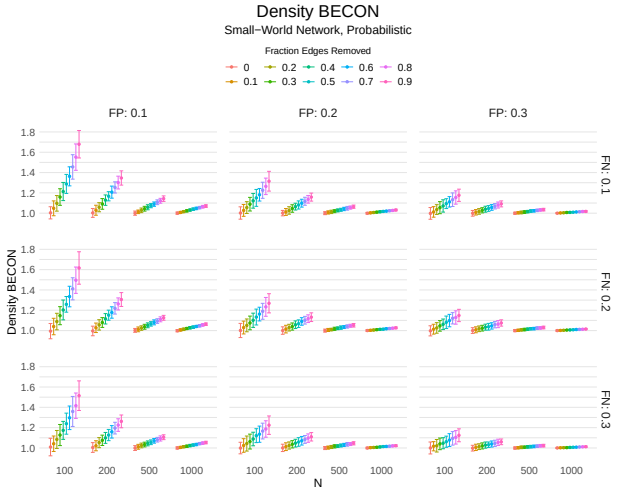


Figure 7. Density BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a small-world network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

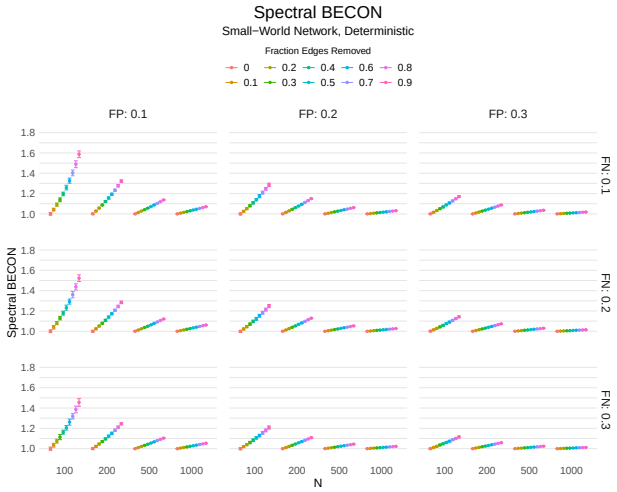


Figure 8. Spectral BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a small-world network. A fixed number of false positives and negatives were added in each run to create the observed network.

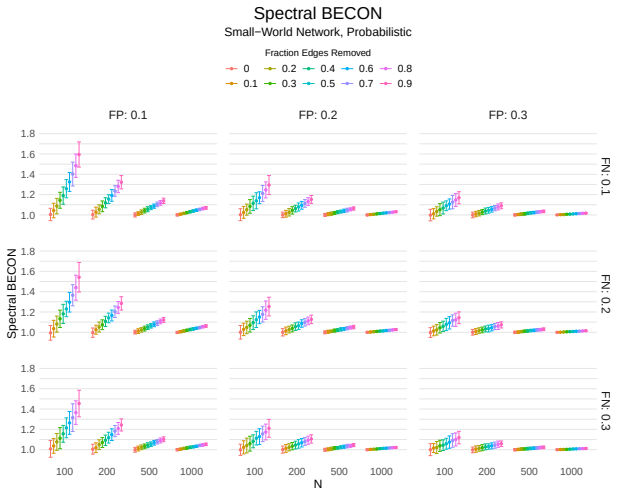


Figure 9. Spectral BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a small-world network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

A.2 Scale-free graph

Table 2. Simulation results across conditions for a scale free graph. The table reports correlations (means and sd) between BECONs and intervention effect sizes for deterministic and probabilistic interventions across simulation runs for multiple network sizes. The results are averaged over the false negative rates.

n	FP	Deterministic						Probabilistic					
		Density		Spectral		ASPL		Density		Spectral		ASPL	
		r	SD	r	SD	r	SD	r	SD	r	SD	r	SD
100	0.1	1	0	1	0	1	0	0.97	0.02	0.98	0.02	0.97	0.02
100	0.2	1	0	1	0	1	0	0.92	0.06	0.93	0.06	0.92	0.07
100	0.3	1	0	1	0	1	0	0.86	0.1	0.87	0.1	0.85	0.1
200	0.1	1	0	1	0	1	0	0.97	0.03	0.98	0.02	0.97	0.03
200	0.2	1	0	1	0	1	0	0.92	0.06	0.93	0.06	0.91	0.06
200	0.3	1	0	1	0	1	0	0.85	0.1	0.86	0.1	0.84	0.11
500	0.1	1	0	1	0	1	0	0.97	0.03	0.97	0.02	0.97	0.03
500	0.2	1	0	1	0	1	0	0.92	0.06	0.92	0.05	0.92	0.06
500	0.3	1	0	1	0	1	0	0.86	0.1	0.86	0.09	0.86	0.1
1000	0.1	1	0	1	0	1	0	0.97	0.03	0.97	0.02	0.97	0.03
1000	0.2	1	0	1	0	1	0	0.92	0.06	0.92	0.05	0.92	0.06
1000	0.3	1	0	1	0	1	0	0.86	0.1	0.86	0.1	0.86	0.1

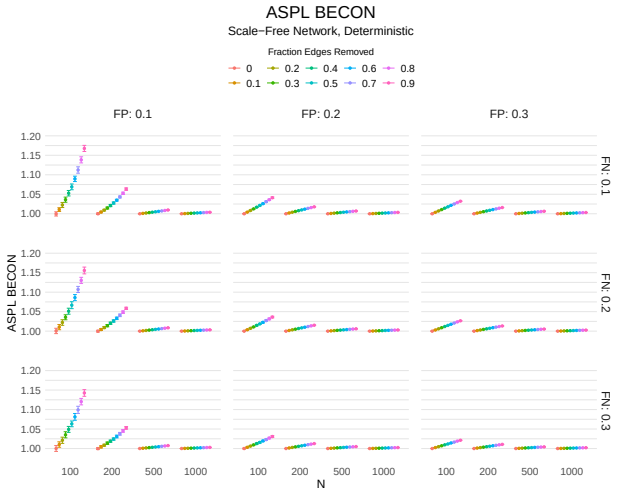


Figure 10. ASPL BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a scale free network. A fixed number of false positives and negatives were added in each run to create the observed network

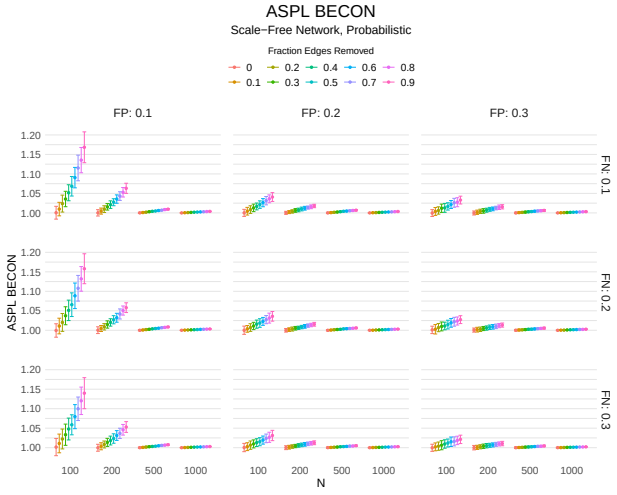


Figure 11. ASPL BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a scale free network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

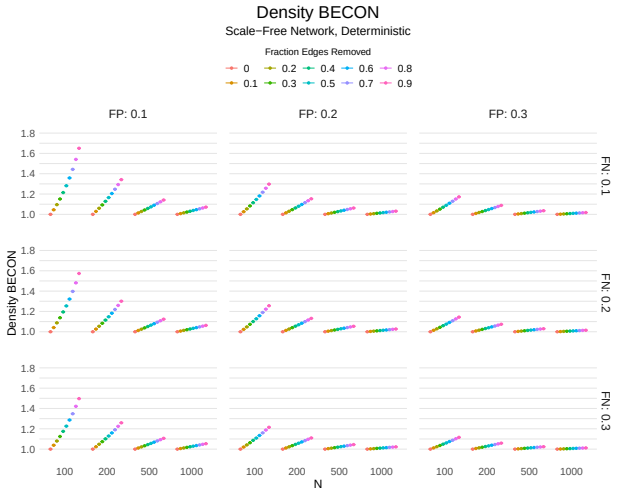


Figure 12. Density BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a scale free network. A fixed number of false positives and negatives were added in each run to create the observed network.

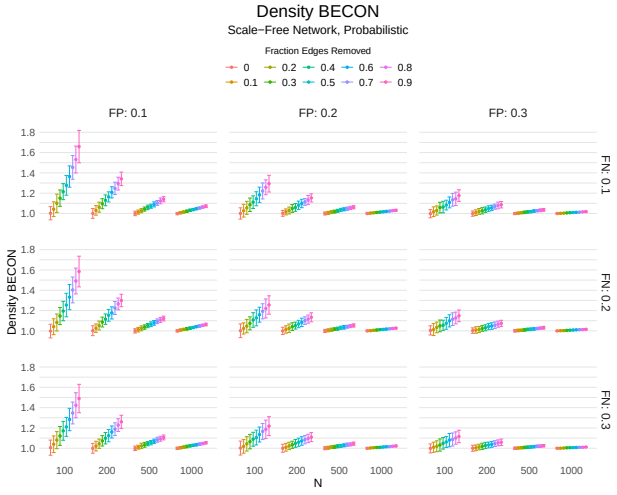


Figure 13. Density BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a scale free network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

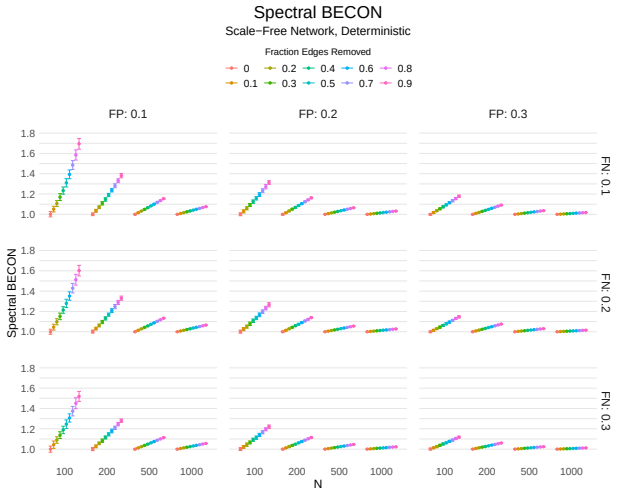


Figure 14. Spectral BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a scale free network. A fixed number of false positives and negatives were added in each run to create the observed network.

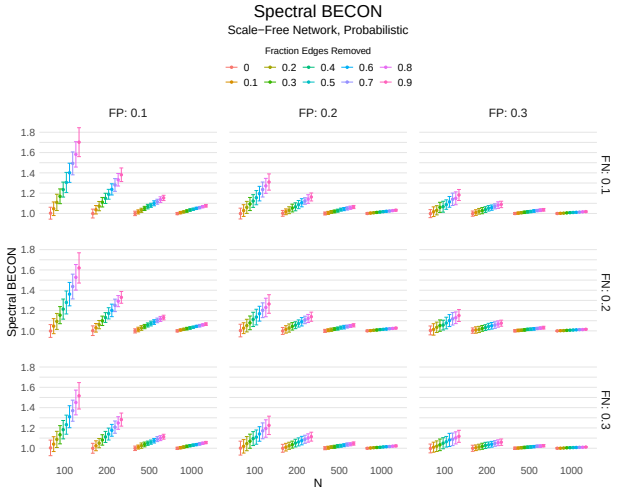


Figure 15. Spectral BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is a scale free network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

A.3 Erdős–Rényi random graph

Table 3. Simulation results across conditions for a Erdős–Rényi graph. The table reports correlations (means and sd) between BECONs and intervention effect sizes for deterministic and probabilistic interventions across simulation runs for multiple network sizes. The results are averaged over the false negative rates.

n	FP	Deterministic						Probabilistic					
		Density		Spectral		ASPL		Density		Spectral		ASPL	
		r	SD	r	SD	r	SD	r	SD	r	SD	r	SD
100	0.1	1	0	1	0	1	0	0.97	0.02	0.97	0.02	0.97	0.02
100	0.2	1	0	1	0	1	0	0.92	0.06	0.92	0.06	0.92	0.06
100	0.3	1	0	1	0	1	0	0.85	0.11	0.85	0.11	0.84	0.11
200	0.1	1	0	1	0	1	0	0.97	0.02	0.97	0.03	0.97	0.02
200	0.2	1	0	1	0	1	0	0.92	0.05	0.92	0.05	0.92	0.06
200	0.3	1	0	1	0	1	0	0.85	0.09	0.85	0.1	0.85	0.09
500	0.1	1	0	1	0	1	0	0.97	0.03	0.97	0.03	0.97	0.03
500	0.2	1	0	1	0	1	0	0.92	0.06	0.92	0.06	0.92	0.06
500	0.3	1	0	1	0	1	0	0.85	0.1	0.85	0.1	0.85	0.1
1000	0.1	1	0	1	0	1	0	0.97	0.03	0.97	0.03	0.96	0.03
1000	0.2	1	0	1	0	1	0	0.91	0.06	0.91	0.06	0.91	0.06
1000	0.3	1	0	1	0	1	0	0.85	0.11	0.85	0.11	0.85	0.11

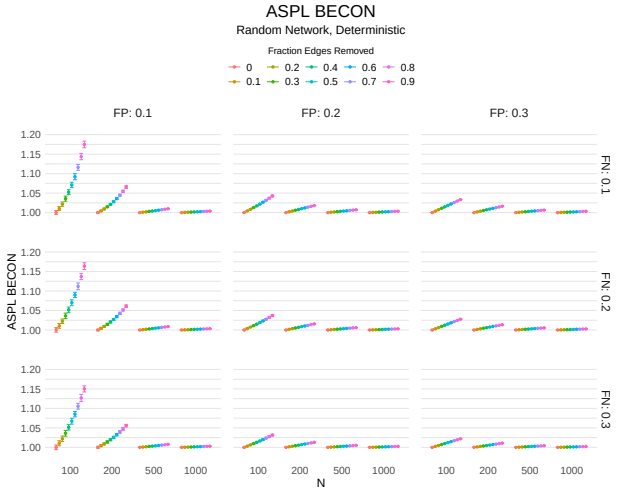


Figure 16. ASPL BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is an Erdős–Rényi random network. A fixed number of false positives and negatives were added in each run to create the observed network

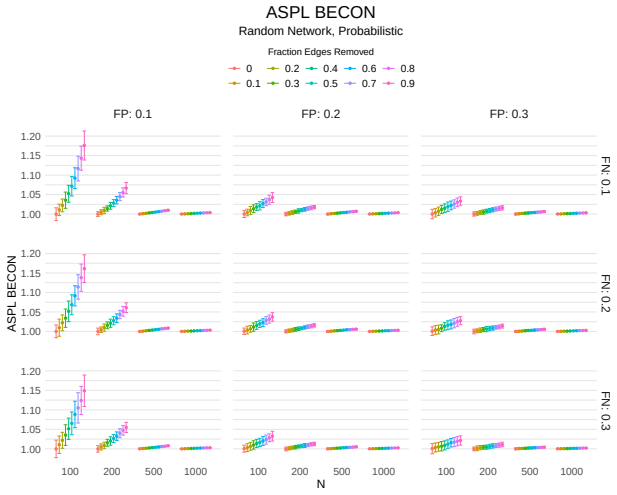


Figure 17. ASPL BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is an Erdős–Rényi random network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

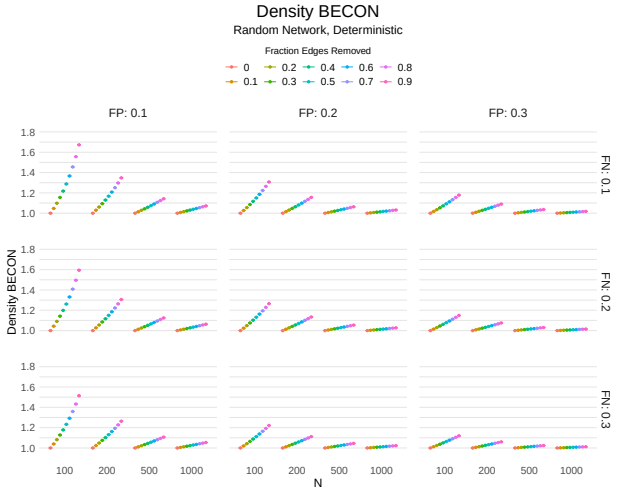


Figure 18. Density BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is an Erdős–Rényi random network. A fixed number of false positives and negatives were added in each run to create the observed network.

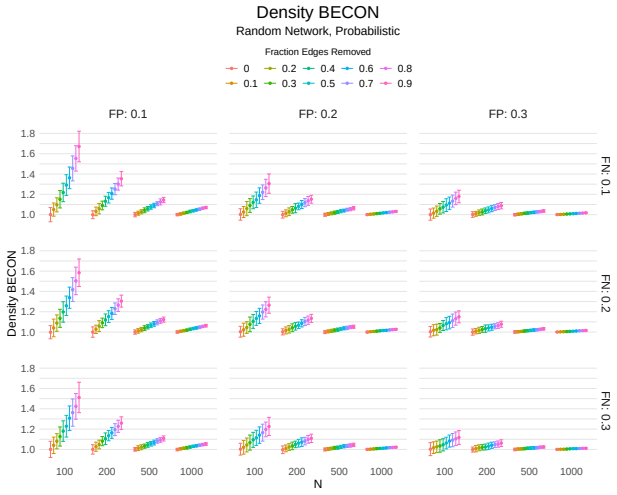


Figure 19. Density BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is an Erdős–Rényi random network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

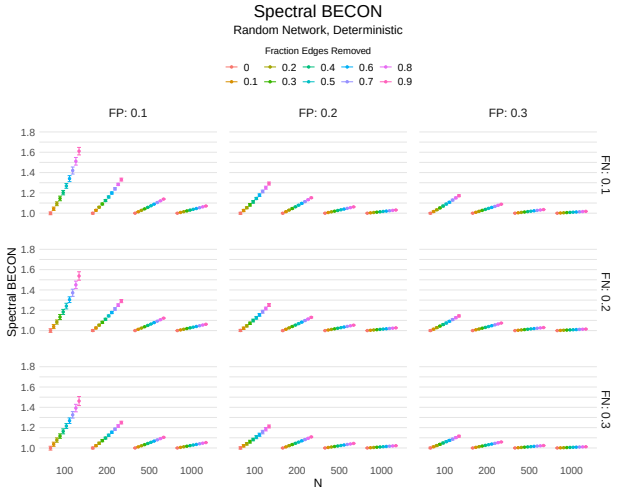


Figure 20. Spectral BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is an Erdős–Rényi random network. A fixed number of false positives and negatives were added in each run to create the observed network.

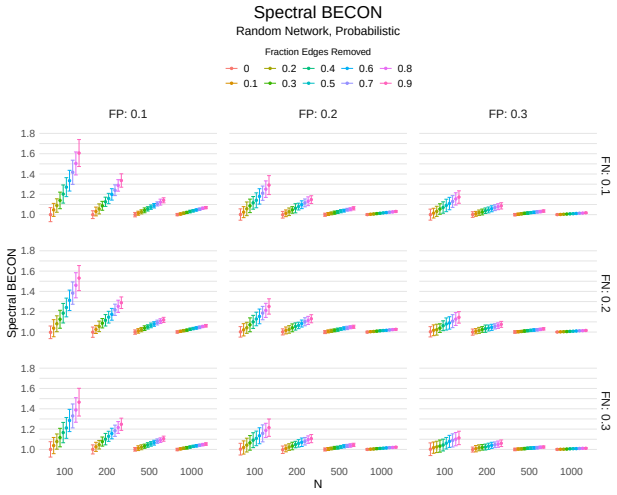


Figure 21. Spectral BECON indicator values as a function of the network size, the intervention effect sizes, the false positive rate, and the false negative rate. The true network structure is an Erdős–Rényi random network. In the observed network, each absent edge in the true network had a probability, FP, of becoming a false positive of and each present edge a probability, FN, of becoming a false negative.

Journal of Behavioral Data Science, 2022, *2* (1), 35–58.
DOI: https://doi.org/10.35566/jbds/v2n1/p2

How to Select the Best Fit Model Among Bayesian Latent Growth Models for Complex Data

Zhenqiu (Laura) Lu*[1][0000−0001−9482−1368] and Zhiyong Zhang[2][0000−0003−0590−2196]

[1] University of Georgia, Athens, GA 30602, USA
zlu@uga.edu
[2] University of Notre Dame, Notre Dame, IN 46530, USA
zhiyongzhang@nd.edu

Abstract. Bayesian approach is becoming increasingly important as it provides many advantages in dealing with complex data. However, there is no well-defined model selection criterion or index in a Bayesian context. To address the challenges, new indices are needed. The goal of this study is to propose new model selection indices and to investigate their performances in the framework of latent growth mixture models with missing data and outliers in a Bayesian context. We consider latent growth models because they are very flexible in modeling complex data and becoming increasingly popular in statistical, psychological, behavioral, and educational areas. Specifically, this study conducted five simulation studies to cover different cases, including latent growth curve models with missing data, latent growth curve models with missing data and outliers, growth mixture models with missing data and outliers, extended growth mixture models with missing data and outliers, and latent growth models with different classes. Simulation results show that almost all proposed indices can effectively identify the true model. This study also illustrated the application of these model selection indices in real data analysis.

Keywords: Model selection criterion · Bayesian estimation · Latent growth models · Missing data · Robust method

1 Introduction

Bayesian approach is becoming increasingly important in estimating models as it provides many advantages in dealing with complex data (e.g., Dunson, 2000). However, there is no well-defined model selection criterion or index in a Bayesian context (e.g., Celeux, Forbes, Robert, & Titterington, 2006). It is due to at least three problems. First, in a Bayesian context there are two versions of deviance

because the Bayesian procedure generates Monte Carlo Markov chains for each parameter. One version is the posterior estimate, which can be estimated by a function of an estimate of a parameter. Another version is the Monte Carlo estimate of the expected deviance based on Bayesian iterations, which can be estimated as the posterior mean of a converged Markov chain. In short, the former is the deviance of the averaged estimates, and the latter is the average of all deviance iterations. The second problem is related to the complexity of the raw data. The data often come from heterogeneous populations which almost unavoidable contain outliers and missing values. The estimates from mis-specified models may result in severely misleading conclusions. The third problem relates to the likelihood function. When latent variables are considered in statistical models, the likelihood function can be an observed-data likelihood function, a complete-data likelihood function, or a conditional likelihood function (Celeux et al., 2006). Furthermore, if data come from heterogeneous populations, the class membership indicator may have different versions, for example, a posterior mode or a posterior mean. Also, with missing data, the likelihood functions have different ways to construct.

1.1 Model Selection Criteria/Indices

Traditional model selection criteria or indices are available for researchers who try to select the best-fit model from a large set of candidate models. Akaike (1974) proposed the Akaike's information criterion (AIC), which offers a relative measure of the information lost. For Bayesian models the Bayes factor, which is the ratio of posterior odds to prior odds, can work for both hypothesis testing and model comparison. But the Bayes factor is often difficult or impossible to calculate, especially for models that involve random effects, large numbers of unknowns or improper priors. To approximate the Bayes factor, Schwarz (1978) developed the Bayesian information criterion (BIC, sometimes called the Schwarz criterion). To obtain more precise indices, Bozdogan (1987) proposed the consistent Akaike information criterion (CAlC), and Sclove (1987) proposed the sample-size adjusted Bayesian information criterion (ssBIC). The deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Linde, 2002) is a recently developed criterion designed for hierarchical models. It is based on the posterior distribution of the log-likelihood and is useful in Bayesian model selection problems where the posterior distributions have been obtained by Markov chain Monte Carlo (MCMC) simulation. DIC is usually regarded as a generalization of AIC and BIC. It is defined analogously to AIC or BIC with a penalty term of the number equal to effective model parameters in Bayesian models. In practice, rough DIC (RDIC or DICV in some literature, e.g., Oldmeadow & Keith, 2011) is an approximation of DIC. The mathematical forms of AIC, BIC, CAIC, ssBIC, and DIC are closely related to each other. They all try to find a balance between the accuracy and the complexity of the fitting model. For all indices above, the model with a smaller criterion/index value is better supported by data.

Lu, Zhang, and Cohen (2013) proposed a series of Bayesian model selection indices based on the traditional ones. However, in Lu et al. (2013) the performances of these indices were investigated when data were non-mixture, normally distributed, and with simple non-ignorable missingness. And only latent growth models were used.

1.2 Goals and Structure

To address the challenges in model selection criterion/index in a Bayesian context, this paper proposes ten model selection indices. This paper also examines the performance of these indices under various conditions by conducting five simulation studies to cover different latent growth models, such as the robust growth models for non-normally distributed data, robust growth mixture models, and the extended robust growth mixture models with missing values. We consider latent growth models because they are very flexible in modeling complex data and becoming increasingly popular in statistical, psychological, behavioral, and educational areas.

The rest of the article consists of five sections. Section 2 presents and formulates three types of models we used in this paper: latent growth models (including growth curve models, growth mixture models, and extended growth mixture models), robust growth models (including three types of robust models), and models that account for missingness (we focus on non-ignorable missingness). Section 3 proposes ten model selection indices in the framework of Bayesian growth models with missing data. Section 4 conducts five simulation studies to evaluate the performance of the Bayesian indices. Model selection results are analyzed, summarized, and compared. Section 5 illustrates the application of these model selection indices in real data analysis. Section 6 discusses the implications and future directions of this study.

2 Latent Growth Models, Robust Growth Models, and Missing Values

Our investigation of the performance of the Bayesian selection indices involves fitting growth models to complex data. In this section, different types of growth models are briefly introduced. Given the fact that the data used in growth models are almost inevitably contain attrition (e.g., Little & Rubin, 2002; Lu, Zhang, & Lubke, 2011; Yuan & Lu, 2008) and outliers (e.g., Maronna, Martin, & Yohai, 2006), different types of growth models are developed, which include traditional latent growth curve models with missing data (Lu et al., 2013), robust growth curve models (Zhang, Lai, Lu, & Tong, 2013) with missing data (Lu & Zhang, 2021), growth mixture models (e.g., Bartholomew & Knott, 1999) with missing data (Lu & Zhang, 2014), extended growth mixture models (EGMMs, Muthén & Shedden, 1999) with missing data (Lu & Zhang, 2014), and robust growth mixture models with missing data (Lu & Zhang, 2014).

In the following, we discuss three types of models: latent growth models (including growth curve models, growth mixture models, and extended growth mixture models), robust growth models (including three types of robust models), and models that account for missingness (we focus on non-ignorable missingness). By combining different elements of these models, it becomes possible to consider a series of growth models with a variety of missing data mechanisms and contaminated data.

2.1 Latent Growth Models

The mathematical form of a latent growth curve model is

$$\begin{cases} \mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \mathbf{e}_i \\ \boldsymbol{\eta}_i = \boldsymbol{\beta} + \boldsymbol{\xi}_i \end{cases}, \tag{1}$$

where $\mathbf{y}_i$ is a $T \times 1$ vector of outcomes for participant $i\,(i = 1, ..., N)$, $\boldsymbol{\eta}_i$ is a $q \times 1$ vector of latent effects, $\boldsymbol{\Lambda}$ is a $T \times q$ matrix of factor loadings for $\boldsymbol{\eta}_i$, $\mathbf{e}_i$ is a $T \times 1$ vector of residual or measurement errors, $\boldsymbol{\beta}$ is a $q \times 1$ vector of fixed-effects, and $\boldsymbol{\xi}_i$ captures the variation of $\boldsymbol{\eta}_i$. We have to note that $\mathbf{e}_i$ and $\boldsymbol{\xi}_i$ are usually assumed normally distributed but not necessary. When data have outliers and are heavy-tailed, this assumption might cause estimate biases. To reduce the effects of outliers, we consider robust models in this study.

A growth mixture model can be expressed as

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \, f_k(\mathbf{y}_i), \tag{2}$$

where π_k is the invariant class probability (or weight) for class k satisfying $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$ (e.g., McLachlan & Peel, 2000), and $f_k(\mathbf{y}_i)(k = 1, \ldots, K)$ is the density of a latent growth model for class k.

For extended growth mixture models (EGMMs, Muthén & Shedden, 1999), π_k is not invariant across individuals. It is allowed to vary individually depending on covariates, so it is expressed as $\pi_{ik}(\mathbf{x}_i)$. If a probit link function is used, then

$$\begin{cases} \pi_{i1}(\mathbf{x}_i) = \Phi(X_i' \, \boldsymbol{\varphi}_1) \\ \pi_{ik}(\mathbf{x}_i) = \Phi(X_i' \, \boldsymbol{\varphi}_k) - \Phi(X_i' \, \boldsymbol{\varphi}_{k-1}), \, (k = 2, 3, ..., K-1), \\ \pi_{iK}(\mathbf{x}_i) = 1 - \Phi(X_i' \, \boldsymbol{\varphi}_{K-1}) \end{cases} \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, and $X_i = (1, \mathbf{x}_i')'$ with an $r \times 1$ vector of observed covariates $\mathbf{x}_i$. Note that $\Phi(X_i' \, \boldsymbol{\varphi}_k) = \sum_{j=1}^{k} \pi_{ij}(\mathbf{x}_i)$ and $\Phi(X_i' \, \boldsymbol{\varphi}_K) \equiv 1$.

A dummy variable $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{iK})'$ is used to indicate the class membership. If individual i comes from group k, $z_{ik} = 1$ and $z_{ij} = 0$ $(\forall j \neq k)$. $\mathbf{z}_i$ is multinomially distributed (McLachlan & Peel, 2000, p.7), that is, $\mathbf{z}_i \sim MultiNomial(\pi_{i1}, \pi_{i2}, ..., \pi_{iK})$.

2.2 Robust Growth Models

When data have outliers and are heavy-tailed, robust methods are used to reduce the effects of outliers. As t-distributions are more robust than normal distributions, the following are robust growth models (Lu & Zhang, 2021; Zhang et al., 2013).

(1) t-Normal (TN) model in which the measurement errors are t-distributed and the latent random effects are normally distributed,

$$\begin{cases} \mathbf{e}_i \sim Mt_T(\mathbf{0}, \boldsymbol{\Theta}, \nu) \\ \boldsymbol{\xi}_i \sim MN_q(\mathbf{0}, \boldsymbol{\Psi}) \end{cases}, \tag{4}$$

where $Mt_T(\mathbf{0}, \boldsymbol{\Theta}, \nu)$ is a T-dimensional multivariate t-distribution with a scale matrix $\boldsymbol{\Theta}$ and degrees of freedom ν, and $MN_q(\mathbf{0}, \boldsymbol{\Psi})$ is a q-dimensional multivariate Normal distribution with a covariance matrix $\boldsymbol{\Psi}$.

(2) Normal-t (NT) model in which the measurement errors are normally distributed but the latent random effects are t-distributed,

$$\begin{cases} \mathbf{e}_i \sim MN_T(\mathbf{0}, \boldsymbol{\Theta}) \\ \boldsymbol{\xi}_i \sim Mt_q(\mathbf{0}, \boldsymbol{\Psi}, u) \end{cases}. \tag{5}$$

(3) t-t (TT) model in which both the measurement errors and the latent random effects are t-distributed,

$$\begin{cases} \mathbf{e}_i \sim Mt_T(\mathbf{0}, \boldsymbol{\Theta}, \nu) \\ \boldsymbol{\xi}_i \sim Mt_q(\mathbf{0}, \boldsymbol{\Psi}, u) \end{cases}. \tag{6}$$

2.3 Missing Values

We focus on the non-ignorable missingness in this paper. To build models with non-ignorable missingness, selection models (Glynn, Laird, & Rubin, 1986; Little, 1993, 1995) are used. For individual i, let $\mathbf{m}_i = (m_{i1}, m_{i2}, ..., m_{iT})'$ be a missing data indicator for $\mathbf{y}_i$, with $m_{it} = 1$ when y_{it} is missing and 0 when observed. Let $\tau_{it} = p(m_{it} = 1)$ be the probability that y_{it} is missing. Then $m_{it} \sim \text{Bernoulli}(\tau_{it})$, so its density function is $f(m_{it}) = \tau_{it}^{m_{it}}(1 - \tau_{it})^{(1-m_{it})}$. The missingness probability τ_{it} can have different forms. Lu and Zhang (2014) proposed the following non-ignorable missingness mechanisms for mixture models.

(1) Latent-Class-Intercept-Dependent (LCID) missingness in which τ_{it} is a function of latent class, covariates, and latent individual initial levels. For example, students are more likely to miss a test if their starting levels of that course are low. We model it as follows:

$$\tau_{it} = \Phi(\mathbf{z}_i'\boldsymbol{\gamma}_{zt} + I_i\gamma_{It} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt}), \tag{7}$$

where I_i is the latent initial levels for individual i, γ_{It} is the coefficient for I_i, $\boldsymbol{\gamma}_{zt}$ is the coefficient for class membership, and $\boldsymbol{\gamma}_{xt}$ are coefficients for covariates. For non-mixture homogenous growth models, LCID can be simplified

to Latent-Intercept-Dependent (LID) without the class membership indicator $\mathbf{z}_i$ and expressed as $\tau_{it} = \Phi(\gamma_{0t} + I_i\gamma_{It} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt})$, where γ_{0t} is the intercept.

(2) Latent-Class-Slope-Dependent (LCSD) missingness in which τ_{it} is a function of latent class, covariates, and latent individual slopes of growth. For example, students are more likely to miss a test if they have slow growth of the course. In this case, τ_{it} can be modeled as

$$\tau_{it} = \Phi(\mathbf{z}_i'\boldsymbol{\gamma}_{zt} + S_i\gamma_{St} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt}), \tag{8}$$

where S_i is the latent slope for individual i, and γ_{St} is the coefficient for S_i. Similarly, for non-mixture homogeneous growth models, LCSD is simplified to Latent-Slope-Dependent (LSD) case as $\tau_{it} = \Phi(\gamma_{0t} + S_i\gamma_{St} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt})$.

(3) Latent-Class-Outcome-Dependent (LCOD) missingness in which τ_{it} is a function of latent class, covariates, and potential outcomes that may be missing. For example, a student who feels he is not doing well on the test may be more likely to give up taking the rest of the test. We express τ_{it} as

$$\tau_{it} = \Phi(\mathbf{z}_i'\boldsymbol{\gamma}_{zt} + y_{it}\gamma_{yt} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt}), \tag{9}$$

where y_{it} is the potential outcomes for individual i at time t, and γ_{yt} is the coefficient for y_{it}. And LCOD can be simplified to Latent-Outcome-Dependent (LOD) for non-mixture homogeneous growth models with a probability of missingness $\tau_{it} = \Phi(\gamma_{0t} + y_{it}\gamma_{yt} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt})$.

In a more general framework, LCID and LCSD can be further encompassed into Latent-Class-Random Effect-Dependent missingness as intercept and slope are different random effects according to different situations under consideration. And for non-mixture structure, LID and LSD are encompassed into Latent-Random Effect-Dependent missingness.

3 Bayesian Model Selection Indices

In this section, we propose ten model selection criteria in the framework of Bayesian growth models with missing data. The definitions of the selection criteria are listed in Table 1. The model selection criteria in the table are based on two versions of deviance in the Bayesian context, $E_{D|y}[D(\theta)]$ and $D(E_{\theta|y}[\theta])$. As we have discussed in the introduction section, $E_{\theta|y}[D]$ is the expected value of all the deviances, and $D(E_{\theta|y}[\theta])$ is the deviance score based on the expected parameters. For different models, the detailed mathematical forms of these two deviances are different. In this paper, we focus on both homogeneous and heterogeneous latent growth models with non-ignorable missing data.

We first look at the homogeneous growth curve models with non-ignorable missing data. One version of deviance, $E_{D|y}[D(\theta)]$, is approximated by

$$\begin{aligned} E_{D|y}[D(\theta)] \approx \overline{D(\theta)} &= -\frac{2}{S}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{t=1}^{T} l_{it}^{(s)}(\theta|y,m) \\ &= -\frac{2}{S}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[(1-m_{it}^{(s)})l_{it}^{(s)}(y) + l_{it}^{(s)}(m)\right], \end{aligned} \tag{10}$$

Table 1. Model Selection Indices

Index =	Deviance + Penalty	
Dbar.AIC[1]	$\overline{D(\theta)}$[4]	2 p
Dbar.BIC[2]	$\overline{D(\theta)}$	log(N) p
Dbar.CAIC	$\overline{D(\theta)}$	(log(N)+1) p
Dbar.ssBIC	$\overline{D(\theta)}$	log((N+2)/24) p
RDIC	$\overline{D(\theta)}$	var(Dbar)/2
Dhat.AIC	$D(\hat{\theta})$[5]	2 p
Dhat.BIC	$D(\hat{\theta})$	log(N) p
Dhat.CAIC	$D(\hat{\theta})$	(log(N)+1) p
Dhat.ssBIC	$D(\hat{\theta})$	log((N+2)/24) p
DIC[3]	$D(\hat{\theta})$	2 pD

Note.

1. p is the number of parameters.
2. N is the sample size.
3. $pD = \overline{D(\theta)} - D(\hat{\theta})$.
4. $\overline{D(\theta)}$ is shown as in eqn.(10) for growth curve models and as in eqn.(13) for growth mixture models.
5. $D(\hat{\theta})$ is shown as in eqn.(12) for growth curve models and as in eqn.(14) for growth mixture models.

where S is the number of iterations for converged Markov chains, $l_{it}^{(s)}(\theta|y,m) = log(L_{it}^{(s)}(\theta|y,m))$ is a conditional joint loglikelihood function (see, Celeux et al., 2006) of y and m, m_{it} is the missing data indicator for individual i at time t with a likelihood function $l_{ikt}(m) = m_{it}log(\tau_{it}) + (1 - m_{it})log(1 - \tau_{it})$, where τ_{it} is the missing data rate for individual i at time t and is defined differently for different missingness models as in the previous section. When y_{it} is missing, the corresponding likelihood is excluded. So combining y and m, the conditional likelihood function of a selection model with non-ignorable missing data can be expressed as

$$L_{it}(\theta|y,m) = [f(y_{it}|\boldsymbol{\eta}_i)(1-\tau_{it})]^{(1-m_{it})}\ \tau_{it}^{m_{it}}, \tag{11}$$

And the other version of deviance, $D(E_{\theta|y}[\theta])$, is approximated by

$$D(E_{\theta|y}[\theta]) \approx D(\hat{\theta}) = -2\sum_{i=1}^{N}\sum_{t=1}^{T}\left[(1-m_{it})l_{it}(y|\hat{\theta}) + l_{it}(m|\hat{\theta})\right], \tag{12}$$

where $\hat{\theta}$ is the posterior mean of parameter estimates across S iterations.

For growth mixture models with missing data, $E_{\theta|y}[D]$ is expressed as

$$E_{D|y}[D(\theta)] \approx \overline{D(\theta)} = -\frac{2}{S}\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}^{(s)}\sum_{t=1}^{T}\left[(1-m_{it})l_{ikt}^{(s)}(y) + l_{ikt}^{(s)}(m)\right], \tag{13}$$

where $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{iK})$ is the class membership indicator which follows a multinomial distribution, $\mathbf{z}_i \sim MultiNomial(\pi_{i1}, \pi_{i2}, ..., \pi_{iK})$, and $z_{ik}^{(s)}$ is the class membership estimated at iteration s. And

$$D(E_{\theta|y}[\theta]) \approx D(\hat{\theta}) = -2 \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{z}_{ik} \sum_{t=1}^{T} \left[(1 - m_{it}) l_{ikt}(y|\hat{\theta}) + l_{ikt}(m|\hat{\theta}) \right], \quad (14)$$

where $\hat{z}_{ik}$ is the posterior mode of class membership, $\hat{\theta}$ is the posterior mean of parameter estimates across all S iterations. In both $\overline{D(\theta)}$ and $D(\hat{\theta})$ definitions of deviance, $l_{ikt}(y)$ and $l_{ikt}(m)$ are the conditional loglikelihood functions for y_{it} and m_{it}, respectively, for individual i in class k at time t.

The difference between $\overline{D(\theta)}$ and $D(\hat{\theta})$ can be quantified by a statistic called pD (Spiegelhalter et al., 2002),

$$pD = \overline{D(\theta)} - D(\hat{\theta}). \quad (15)$$

Based on the Jensen's inequality (Casella & George, 1992), when $D(\theta)$ is convex, then $\overline{D(\theta)} \geq D(\hat{\theta})$ and as a result pD is positive. When $D(\theta)$ is concave, then $\overline{D(\theta)} \leq D(\hat{\theta})$ and pD is negative.

4 Simulation Studies

In this section, five simulation studies are conducted to evaluate the performance of the Bayesian indices. For each study, four waves of complete data are generated first and then missing data are created on each occasion according to pre-designed missing data rates. After data are generated, full Bayesian methods are used by adopting uninformative priors, obtaining conditional posterior distributions through application of a data augmentation algorithm, generating Markov chains through a Gibbs sampling procedure, conducting convergence testing, and making statistical inference for model parameters. For all simulations, the software OpenBUGS is used for the implementation of Gibbs sampling, and R is used for data-generation, convergence testing, and parameter estimation.

The five studies are designed such that the data complexity increases from study 1 to study 5. Studies 1-2 focus on non-mixture growth data and thus, latent growth curve models with missing data are used. Studies 3-5 focus on mixture growth data and thus, growth mixture models with missing data are used. Simulation factors include measurement error distributions, random-effects distributions, missingness patterns, sample size, and class separation (Anderson & Bahadur, 1962). Under each condition, 100 converged replications are used to calculate the model selection proportion. Table 2 lists the design details.

Table 2: Simulation Study Design

Study	Model	Distribution				Missingness					Sample Size	Class	
		$\mathbf{e}_i$ [2]		$\boldsymbol{\eta}_i$ [3]		Depends on						Separation[11]	
		N[4]	t[5]	N	t	C[6]	X[7]	I[8]	S[9]	Y[10]	Different	M	S
Study 1	Normal LGCMs: use relative small sample sizes due to single-class data												
	NN-ignorable	✓		✓		✓							
	NN-XI	✓		✓		✓	✓						
	NN-XS[1]	✓		✓		✓		✓					
	NN-XY	✓		✓		✓			✓				
Study 2	Robust LGCMs: use relative small sample sizes due to single-class data												
	TN-ignorable		✓	✓		✓							
	TN-XI		✓	✓		✓	✓						
	TN-XS		✓	✓		✓		✓					
	TN-XY		✓	✓		✓			✓				
	TT-ignorable		✓		✓	✓							
	TT-XI		✓		✓	✓	✓						
	TT-XS		✓		✓	✓		✓					
	TT-XY		✓		✓	✓			✓				
	NT-ignorable	✓			✓	✓							
	NT-XI	✓			✓	✓	✓						
	NT-XS	✓			✓	✓		✓					
	NT-XY	✓			✓	✓			✓				
	NN-ignorable	✓		✓		✓							
	NN-XI	✓		✓		✓	✓						
	NN-XS	✓		✓		✓		✓					
	NN-XY	✓		✓		✓			✓				
Study 3	Robust GMMs (RGMMs): use relative large sample sizes due to multiple classes data, and use small class separation due to fixed class probabilities												
	TN-ignorable		✓	✓		✓						✓	
	TN-XI		✓	✓		✓	✓					✓	
	TN-XS		✓	✓		✓		✓				✓	
	TN-XY		✓	✓		✓			✓			✓	
	TT-ignorable		✓		✓	✓						✓	
	TT-XI		✓		✓	✓	✓					✓	
	TT-XS		✓		✓	✓		✓				✓	
	TT-XY		✓		✓	✓			✓			✓	
	NT-ignorable	✓			✓	✓						✓	
	NT-XI	✓			✓	✓	✓					✓	
	NT-XS	✓			✓	✓		✓				✓	

NT-XY	✓		✓		✓		✓		✓	
NN-ignorable	✓	✓			✓				✓	
NN-XI	✓	✓			✓	✓			✓	
NN-XS	✓	✓			✓		✓		✓	
NN-XY	✓	✓			✓			✓	✓	
Study 4 Robust Extended GMMs (REGMMs): select 5 competing models based on the performance in Study 3 use relative large sample sizes due to multiple-class data and varied class probabilities										
TN-CXS		✓	✓		✓	✓		✓	✓	✓
TN-CX		✓	✓		✓	✓			✓	✓
TT-CXS		✓		✓	✓	✓		✓	✓	✓
NN-CXS	✓		✓		✓	✓		✓	✓	✓
NN-CX	✓		✓		✓	✓			✓	✓
Study 5 Single-Class LGCMs vs. Multiple-Class RGMMs										
1 Class LGCMs										
TN-XS		✓	✓			✓	✓			
TT-XS		✓		✓		✓	✓			
NN-XS	✓		✓			✓	✓			
2 Classes RGMMs										
TN-XS		✓	✓			✓	✓		✓	
TT-XS		✓		✓		✓	✓		✓	
NN-XS	✓		✓			✓	✓		✓	
3 Classes RGMMs										
TN-XS		✓	✓			✓	✓			
TT-XS		✓		✓		✓	✓			
NN-XS	✓		✓			✓	✓			
4 Classes RGMMs										
TN-XS		✓	✓			✓	✓			
TT-XS		✓		✓		✓	✓			
NN-XS	✓		✓			✓	✓			

Note. 1 The shaded model is the true model. 2 Measurement errors. 3 Random effects. 4 Normal distribution. 5 t distribution. 6 Latent class dependent (Non-ignorable). 7 Observed Covariates. 8 Latent intercept dependent (Non-ignorable). 9 Latent slope dependent (Non-ignorable). 10 Potential outcome y dependent (Non-ignorable). 11 Class Separation (Anderson & Bahadur, 1962) when generating data (S: small=1.7, M: medium=2.7).

Study 1 investigated the performance of the Bayesian indices when data were non-mixture, homogeneous, normally distributed with non-ignorable missingness. The true model was NN-XS, which was the model with normally distributed measurement errors ($\mathbf{e}_i$) at level 1 and random effects ($\boldsymbol{\xi}_i$) at level 2, with missingness depending on covariate x and latent slope S. Specifically,

$\mathbf{e}_i \sim MN(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\eta}_i \sim MN_q(\boldsymbol{\beta}, \boldsymbol{\Psi})$ where $\boldsymbol{\beta}$ = (Intercept, Slope) = $(1, 3)$ and $\boldsymbol{\Psi}$ was a 2 by 2 symmetric matrix with $Var(I) = 1$, $Cov(I, S) = 0$, and $Var(S) = 4$. For missingness, the bigger the latent slope was, the higher the missing data rate would be. The missingness probit coefficients were set as $\gamma_0 = (-1, -1, -1, -1)$, $\gamma_x = (-1.5, -1.5, -1.5, -1.5)$, and $\gamma_S = (0.5, 0.5, 0.5, 0.5)$. For example, if a participant had a latent growth slope 3, with a covariate value 1, then his or her missing probability at each wave was $\tau \approx 16\%$; if the slope was 5, with the same covariate value, the missing probability increased to $\tau = 50\%$; but if the slope was 1, then the missing probability decreased to $\tau = 2.3\%$. The covariate x was also generated from a normal distribution, $x \sim N(1, sd = 0.2)$. In study 1, in total there were 16 conditions with 4 missingness mechanisms (XS non-ignorable, XY non-ignorable, XI non-ignorable, and ignorable) combined with 4 levels of sample size (1000, 500, 300, and 200). Table 3 lists the model selection proportions across 100 replications for each of these indices across all conditions in study 1. The largest proportion across 4 missingness models is indicated in the shaded cell for each index. When sample size is relatively large, 1000 or 500, all of the model selection indices, except for the rough DIC (RDIC), correctly identify the true model with 100%. When sample size becomes smaller, 300 or 200, except for the RDIC, all of the model selection indices choose the true model with certainty above 93%. Comparing the indices defined based on Dbar with those defined based on Dhat, one can see that the former performs a little bit better.

Study 2 investigated the performance of these indices when data were non-mixture homogeneous with outliers and non-ignorable missingness. The main difference between study 2 and 1 was that the data for study 2 contain outliers such that they are not normally distributed. So robust growth curve models were used. The true model was TN-XS, which means measurement errors ($\mathbf{e}_i$) at level 1 followed a t-distribution. Specifically, $\mathbf{e}_i$ were generated from a t distribution with 5 degrees of freedom and a scale matrix $\mathbf{I}$, i.e., $\mathbf{e}_i \sim Mt(\mathbf{0}, \mathbf{I}, 5)$. Other settings were kept the same as those in study 1. In this study, totally 32 conditions were considered with 4 data distributions (NN, TN, NT, and TT), 4 missingness patterns (XS non-ignorable, XY non-ignorable, XI non-ignorable, and ignorable), and 2 levels of sample size (1000 and 500). Table 4 lists the model selection proportions. The largest proportion across 16 missingness models is indicated in the shaded cells for each index. Except for the RDIC, all of the model selection indices correctly identify the true model. TT-XS is a competing model, which also gains high selection probabilities. This is because the normal distribution is almost identical to a t-distribution with large degrees of freedom. The degrees of freedom of t is also estimated by the model. Also, the Dbar-based indices performs a little bit better than the Dhat-based indices. Among them, Dbar-based BIC and CAIC perform best.

Study 3 was designed for mixture data with outliers and non-ignorable missing data. As data were mixture, growth mixture models were used. In this study, the true model was 2-class mixture TN-XS RGMM. Only intercepts of these 2 classes were different, with 5 for class 1 and 1 for class 2. Other

Table 3. Model Selection Proportion in Study 1

	N=1000				N=500			
	Non-ignorable			Ignorable	Non-ignorable			Ignorable
Criteron[1]	NN-XS[2]	NN-XY[3]	NN-XI[4]	NN[5]	NN-XS	NN-XY	NN-XI	NN
Dbar.AIC	1[6]	0.000	0.000	0.000	1	0.000	0.000	0.000
Dbar.BIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
Dbar.CAIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
Dbar.ssBIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
RDIC	0.013	0.000	0.987	0.000	0.038	0.000	0.962	0.000
Dhat.AIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
Dhat.BIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
Dhat.CAIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
Dhat.ssBIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
DIC	1	0.000	0.000	0.000	1	0.000	0.000	0.000
	N=300				N=200			
Dbar.AIC	0.98125	0.01875	0.000	0.000	0.975	0.025	0.000	0.000
Dbar.BIC	0.98125	0.01875	0.000	0.000	0.975	0.025	0.000	0.000
Dbar.CAIC	0.98125	0.01875	0.000	0.000	0.975	0.025	0.000	0.000
Dbar.ssBIC	0.98125	0.01875	0.000	0.000	0.975	0.025	0.000	0.000
Rough DIC	0.1125	0.000	0.8875	0.000	0.2	0.03125	0.76875	0.000
Dhat.AIC	0.95	0.05	0.000	0.000	0.9375	0.06875	0.000	0.000
Dhat.BIC	0.95	0.05	0.000	0.000	0.9375	0.06875	0.000	0.000
Dhat.CAIC	0.95	0.05	0.000	0.000	0.9375	0.06875	0.000	0.000
Dhat.ssBIC	0.95	0.05	0.000	0.000	0.9375	0.06875	0.000	0.000
DIC	1	0.000	0.000	0.000	0.98125	0.0125	0.00625	0.000

Note.

1. The definition of each index is given in Table 1.
2. The shaded model is the true model. The model is normal-distribution-based with
latent-slope-dependent missingness.
3. The model is normal-distribution-based with potential-outcome-dependent missingness.
4. The model is normal-distribution-based with latent-intercept-dependent missingness.
5. The model is normal-distribution-based with ignorable missingness.
6. The shaded cell has the largest proportion.

Table 4. Model Selection Proportion in Study 2

		N=1000				N=500			
		Non-ignorable			Ignorable	Non-ignorable			Ignorable
Index		XS[5]	XY	XI		XS	XY	XI	
Dbar.AIC	TN[1]	0.519	0.000	0.000	0.000	0.597	0.013	0.000	0.000
	TT[2]	0.469	0.000	0.000	0.012	0.377	0.000	0.000	0.000
	NT[3]	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000
	NN[4]	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000
Dbar.BIC	TN	0.781	0.000	0.000	0.000	0.855	0.013	0.000	0.000
	TT	0.200	0.000	0.000	0.019	0.113	0.000	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000
	NN	0.000	0.000	0.000	0.000	0.013	0.000	0.000	0.000
Dbar.CAIC	TN	0.819	0.000	0.000	0.000	0.888	0.012	0.000	0.000
	TT	0.162	0.000	0.000	0.019	0.075	0.000	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	NN	0.000	0.000	0.000	0.000	0.019	0.000	0.000	0.000
Dbar.ssBIC	TN	0.625	0.000	0.000	0.000	0.631	0.012	0.000	0.000
	TT	0.362	0.000	0.000	0.012	0.338	0.000	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000
	NN	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000
RDIC	TN	0.000	0.000	0.106	0.000	0.000	0.000	0.094	0.000
	TT	0.000	0.000	0.100	0.000	0.000	0.000	0.113	0.000
	NT	0.000	0.000	0.394	0.000	0.000	0.000	0.390	0.000
	NN	0.000	0.000	0.400	0.000	0.000	0.000	0.403	0.000
Dhat.AIC	TN	0.544	0.000	0.000	0.000	0.547	0.025	0.000	0.000
	TT	0.506	0.006	0.000	0.000	0.447	0.019	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	NN	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Dhat.BIC	TN	0.675	0.006	0.000	0.000	0.717	0.025	0.000	0.000
	TT	0.319	0.000	0.000	0.000	0.245	0.013	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	NN	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Dhat.CAIC	TN	0.700	0.006	0.000	0.000	0.788	0.025	0.000	0.000
	TT	0.294	0.006	0.000	0.000	0.169	0.012	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	NN	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Dhat.ssBIC	TN	0.575	0.006	0.000	0.000	0.588	0.025	0.000	0.000
	TT	0.419	0.006	0.000	0.000	0.369	0.012	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	NN	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DIC	TN	0.325	0.000	0.000	0.000	0.415	0.006	0.000	0.000
	TT	0.462	0.000	0.000	0.194	0.409	0.000	0.000	0.000
	NT	0.012	0.000	0.000	0.000	0.088	0.000	0.000	0.000
	NN	0.006	0.000	0.000	0.000	0.082	0.000	0.000	0.000

Note. [1−4]T-Normal, T-T, Normal-T, and Normal-Normal measurement errors and random effects. [5]Other abbreviations are as given in Table 3.

settings for each class were the same as in study 2. Both classes have t_5 distributed measurement errors. Based on Anderson and Bahadur (1962), the class separation is around 2.7. In this study, we assumed they are traditional mixture models, i.e., class probabilities were fixed at (50%, 50%) in this study. The same as in study 2, there were 32 conditions considered with 4 data distributions (NN, TN, NT, and TT), 4 missingness patterns (XS non-ignorable, XY non-ignorable, XI non-ignorable, and ignorable), and 2 levels of sample size (1000 and 1500). As mixture data require more data to obtain estimates, we increased the sample size. Table 5 shows the results for study 3. The shaded cell indicates the largest proportion across 16 missingness models for each index. Again, almost all of the model selection indices correctly identify the true model. And the Dbar-based indices perform a little bit better than the Dhat-based indices. Specifically, Dbar-based BIC and CAIC perform best among these indices, and then Dbar-based ssBIC also perform well.

Study 4 extended study 3 such that the class probabilities were not fixed. Instead, they depended on values of covariates. Also, the non-ignorable missingness in this study was allowed to depend on the corresponding observations' latent class membership. The true model in this study was 2-class mixture TN-CXS robust extended growth mixture models (REGMM). The differences between this study and study 3 were (1) the class proportions in this study were predicted by the value of covariate x; (2) the missing data rates were predicted by the latent class membership; and (3) both medium size, 2.7, and small size, 1.7, class separations were used. Specifically, for small class separation, the intercept for class 1 was 3.5 and the intercept for class 2 was 1. To simplify the simulation, based on the findings in study 3, 5 competing mixture models (TN-CXS, TT-CXS, TN-CX, NN-CXS, and NN-CX) were chosen to fit the data. Totally, we considered 20 conditions with 5 mixture models, 2 levels of sample size (1500 and 1000), and 2 levels of class separation (2.7 and 1.7). Table 6 shows the model selection proportions in study 4. Again, almost all of the model selection indices correctly identify the true model. Specifically, Dbar-based BIC and CAIC perform best among these indices.

Study 5 focused on the number of classes. In this study, different growth curve models with different numbers of classes were fitted and compared. In total, 9 conditions were considered, including 3 models (TN-XS, TT-XS, NN-XS) and 3 numbers of classes (1, 2, and 3). The true model was the 2-class mixture TN-XS model. The simulation results for study 5 were presented in Table 7. Among these indices, Dhat-based indices perform better than Dhbar-based indices. Specifically, Dhat-based BIC and CAIC perform best, and ssBIC and AIC also provide high certainty.

Table 5. Model Selection Proportion in Study 3

		N=1500				N=1000			
		Non-ignorable			Ignorable	Non-ignorable			Ignorable
Index		XS	XY	XI		XS	XY	XI	
Dbar.AIC	TN	0.621	0.000	0.000	0.000	0.593	0.000	0.000	0.000
	TT	0.357	0.000	0.000	0.000	0.314	0.000	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.021	0.000	0.000	0.000
	NN	0.021	0.000	0.000	0.000	0.071	0.000	0.000	0.000
Dbar.BIC	TN	0.864	0.000	0.000	0.000	0.843	0.000	0.000	0.000
	TT	0.114	0.000	0.000	0.000	0.064	0.000	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.000
	NN	0.021	0.007	0.000	0.000	0.079	0.000	0.000	0.000
Dbar.CAIC	TN	0.893	0.000	0.000	0.000	0.857	0.000	0.000	0.000
	TT	0.079	0.000	0.000	0.000	0.043	0.000	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.007	0.007	0.000	0.000
	NN	0.021	0.007	0.000	0.000	0.086	0.000	0.000	0.000
Dbar.ssBIC	TN	0.729	0.000	0.000	0.000	0.750	0.000	0.000	0.000
	TT	0.250	0.000	0.000	0.000	0.157	0.000	0.000	0.000
	NT	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.000
	NN	0.021	0.007	0.000	0.000	0.079	0.000	0.000	0.000
RDIC	TN	0.071	0.000	0.000	0.000	0.143	0.000	0.000	0.000
	TT	0.086	0.000	0.000	0.000	0.071	0.000	0.000	0.000
	NT	0.450	0.000	0.000	0.000	0.393	0.007	0.000	0.000
	NN	0.393	0.000	0.000	0.000	0.379	0.007	0.000	0.000
Dhat.AIC	TN	0.586	0.000	0.000	0.000	0.621	0.000	0.000	0.000
	TT	0.379	0.000	0.000	0.000	0.329	0.000	0.000	0.000
	NT	0.014	0.000	0.000	0.000	0.014	0.007	0.000	0.000
	NN	0.014	0.007	0.000	0.000	0.057	0.000	0.000	0.000
Dhat.BIC	TN	0.757	0.000	0.000	0.000	0.793	0.000	0.000	0.000
	TT	0.207	0.000	0.000	0.000	0.121	0.000	0.000	0.000
	NT	0.007	0.000	0.000	0.000	0.007	0.007	0.000	0.000
	NN	0.021	0.007	0.000	0.000	0.071	0.000	0.000	0.000
Dhat.CAIC	TN	0.757	0.000	0.000	0.000	0.814	0.000	0.000	0.000
	TT	0.207	0.000	0.000	0.000	0.100	0.000	0.000	0.000
	NT	0.007	0.000	0.000	0.000	0.007	0.007	0.000	0.000
	NN	0.021	0.007	0.000	0.000	0.071	0.000	0.000	0.000
Dhat.ssBIC	TN	0.586	0.000	0.000	0.000	0.664	0.000	0.000	0.000
	TT	0.379	0.000	0.000	0.000	0.250	0.000	0.000	0.000
	NT	0.014	0.000	0.000	0.000	0.014	0.007	0.000	0.000
	NN	0.014	0.007	0.000	0.000	0.064	0.000	0.000	0.000
DIC	TN	0.507	0.000	0.000	0.000	0.364	0.007	0.000	0.000
	TT	0.371	0.000	0.000	0.000	0.286	0.000	0.000	0.000
	NT	0.043	0.036	0.000	0.000	0.129	0.029	0.007	0.000
	NN	0.043	0.000	0.000	0.000	0.150	0.029	0.000	0.000

Note. Same as Table 3.

Table 6. Model Selection Proportion in Study 4

Index	TN-CXS	TT-CXS	NN-CXS	TN-CX	NN-CX
Class Separation=2.7, N=1500					
Dbar.AIC	0.567	0.425	0.000	0.008	0.000
Dbar.BIC	0.808	0.158	0.000	0.033	0.000
Dbar.CAIC	0.850	0.108	0.000	0.0042	0.000
Dbar.ssBIC	0.667	0.300	0.000	0.033	0.000
RDIC	0.042	0.042	0.908	0.000	0.008
[1pt] Dhat.AIC	0.475	0.392	0.000	0.133	0.000
Dhat.BIC	0.550	0.233	0.000	0.217	0.000
Dhat.CAIC	0.525	0.233	0.000	0.242	0.000
Dhat.ssBIC	0.467	0.367	0.000	0.167	0.000
DIC	0.467	0.500	0.033	0.000	0.000
Class Separation=2.7, N=1000					
Dbar.AIC	0.558	0.375	0.000	0.067	0.000
Dbar.BIC	0.750	0.125	0.000	0.125	0.000
Dbar.CAIC	0.767	0.100	0.008	0.125	0.000
Dbar.ssBIC	0.633	0.292	0.000	0.075	0.000
RDIC	0.092	0.075	0.808	0.000	0.025
[1pt] Dhat.AIC	0.350	0.358	0.000	0.292	0.000
Dhat.BIC	0.450	0.175	0.000	0.375	0.000
Dhat.CAIC	0.442	0.150	0.000	0.4	0.008
Dhat.ssBIC	0.392	0.300	0.000	0.308	0.000
DIC	0.417	0.450	0.108	0.008	0.017
Class Separation=1.7, N=1500					
Dbar.AIC	0.512	0.444	0.044	0.000	0.00
Dbar.BIC	0.744	0.212	0.044	0.000	0.00
Dbar.CAIC	0.781	0.175	0.044	0.000	0.00
Dbar.ssBIC	0.612	0.344	0.044	0.000	0.00
RDIC	0.306	0.238	0.350	0.006	0.10
[1pt] Dhat.AIC	0.475	0.475	0.031	0.019	0.00
Dhat.BIC	0.712	0.238	0.031	0.019	0.00
Dhat.CAIC	0.712	0.238	0.031	0.019	0.00
Dhat.ssBIC	0.475	0.475	0.031	0.019	0.00
DIC	0.381	0.450	0.169	0.000	0.00
Class Separation=1.7, N=1000					
Dbar.AIC	0.550	0.400	0.050	0.000	0.000
Dbar.BIC	0.719	0.194	0.081	0.006	0.000
Dbar.CAIC	0.750	0.162	0.081	0.006	0.000
Dbar.ssBIC	0.638	0.300	0.062	0.000	0.000
RDIC	0.244	0.256	0.362	0.000	0.138
[1pt] Dhat.AIC	0.694	0.231	0.012	0.062	0.000
Dhat.BIC	0.644	0.294	0.012	0.050	0.000
Dhat.CAIC	0.694	0.231	0.012	0.062	0.000
Dhat.ssBIC	0.575	0.388	0.012	0.025	0.000
DIC	0.344	0.331	0.319	0.000	0.006

Note. Same as Table 3.

Table 7. Model Selection Proportion in Study 5

	2 CLASSES			1 CLASS			3 CLASSES		
Index	TN-XS	TT-XS	NN-XS	TN-XS	TT-XS	NN-XS	TN-XS	TT-XS	NN-XS
Dbar.AIC	0.000	0.000	0.057	0.393	0.129	0.000	0.021	0.007	0.393
Dbar.BIC	0.000	0.000	0.036	0.821	0.064	0.000	0.000	0.000	0.079
Dbar.CAIC	0.000	0.000	0.036	0.864	0.043	0.000	0.000	0.000	0.057
Dbar.ssBIC	0.000	0.000	0.057	0.593	0.100	0.000	0.000	0.000	0.25
RDIC	0.036	0.014	0.2	0.014	0.014	0.679	0.014	0.014	0.014
Dhat.AIC	0.621	0.343	0.064	0.000	0.000	0.000	0.000	0.000	0.000
Dhat.BIC	0.793	0.136	0.071	0.000	0.000	0.000	0.000	0.000	0.000
Dhat.CAIC	0.814	0.114	0.071	0.000	0.000	0.000	0.000	0.000	0.000
Dhat.ssBIC	0.664	0.264	0.071	0.000	0.000	0.000	0.000	0.000	0.000
DIC	0.000	0.000	0.000	0.164	0.193	0.121	0.000	0.000	0.521

Note. Same as Table 3.

5 Application

In this section, a real data set on mathematical growth is analyzed to demonstrate the application of the indices. The same sample that has been analyzed in Lu et al. (2011) is used here. It is a mathematical ability growth sample from the NLSY97 survey (Bureau of Labor Statistics, U.S. Department of Labor, 1997), which were collected from $N = 1510$ adolescents yearly from 1997 to 2001 when each adolescent was administered the Peabody Individual Achievement Test (PIAT) Mathematics Assessment to measure their mathematical ability. There are some outliers at all five grades. Lu et al. (2011) conducted a power transformation to normalize the sample and assumed the data are normally distributed without outliers. In this study, however, we use the original non-transformed data with outliers, but robust methods are used. Also, different non-ignorable missingness mechanisms are considered. Overall, the means of mathematical ability increased over time with a roughly linear trend. The missing data rates range from 4.57% to 9.47%, and the raw data show the missing pattern is intermittent. About half of the sample is female.

The analysis is conducted following the steps in Table 8. In step 1, a tentative model (the TT-ignorable model) is fitted to the data. Gender is a covariate. The estimates of degrees of freedom of t for both classes are 2.342 and 3.263 for measurement errors and 75.65 and 50.96 for random effects, which indicates that measurement errors can be better fitted using a t distribution while random effects are approximately normally distributed (i.e., a TN model). And then in step 2, to compare models with different non-ignorable missingness and numbers of classes, 10 models are fitted to the data. During estimation we

Table 8. Steps and Fitting Models in Real Data Analysis

		$\mathbf{e}_i$		$\boldsymbol{\eta}_i$		missingness				
	Model	N	T	N	T	C	X	I	S	Y
Step 1: Fit a tentative 2 classes model, and check the estimated df of t										
	TT-ignorable	✓		✓						
Step 2: Try models with different missingness and number of classes										
	2 Classes RGMMs									
	TN-X	✓		✓			✓			
	TN-XI	✓		✓			✓	✓		
	TN-XS	✓		✓			✓		✓	
	TN-XY	✓		✓			✓			✓
	2 Classes REGMMs									
	TN-CX	✓		✓		✓	✓			
	TN-CXI	✓		✓		✓	✓	✓		
	TN-CXS	✓		✓		✓	✓		✓	
	TN-CXY	✓		✓		✓	✓			✓
	3 Classes GMMs									
	NN-X	✓		✓			✓			
	4 Classes GMMs									
	NN-X	✓		✓			✓			
Step 3: Compare selection indices										
Step 4: Interpret results obtained from the selected model										

Note. Abbreviations are as given in Table 2.

Table 9. Model Selection in Real Data Analysis

	2 CLASSES								3 CLASSES	4 CLASSES
Index[1]	TN-CXS	TN-CXY	TN-CXI	TN-CX	TN-XS	TN-XY	TN-XI	TN-X	NN-X	NN-X
Dbar.AIC	17392	17472	17502	17502	17392	17482	17502	17512	17372	17126
Dbar.BIC	17583.52	17663.52	17693.52	17666.92	17556.92	17646.92	17666.92	17650.32	17536.92	17328.15
Dbar.CAIC	17619.52	17699.52	17729.52	17697.92	17587.92	17677.92	17697.92	17676.32	17567.92	17366.15
Dbar.ssBIC	17469.15	17549.15	17579.15	17568.44	17458.44	17548.44	17568.44	17567.72	17438.44	17207.44
RDIC	22759.24	22704.5	22378.14	22601.28	22562.65	22755.44	22973.52	22520.18	22843.52	23333.2
Dhat.AIC	15192	14942	17482	19822	21922	23622	25722	27352	15872	15716
Dhat.BIC	15383.52	15133.52	17673.52	19986.92	22086.92	23786.92	25886.92	27490.32	16036.92	15918.15
Dhat.CAIC	15419.52	15169.52	17709.52	20017.92	22117.92	23817.92	25917.92	27516.32	16067.92	15956.15
Dhat.ssBIC	15269.15	15019.15	17559.15	19888.44	21988.44	23688.44	25788.44	27407.72	15938.44	15797.44
DIC	19520	19930	17450	15120	12800	11280	9220	7620	18810	18460

Note. [1]The definition of each index is given in Table 1. [2]The shaded cell has the smallest value.

Table 10. Estimates of TN-CXY REGMM in Real Data Analysis

		Parameter	Mean	S.D.	MC.e./S.D.[1]	Lower[2]	Upper[3]	Geweke z[4]
Growth Curve Parameters	Class 1	Intercept	8.647	0.037	0.026	8.572	8.717	0.007
		Slope	0.229	0.009	0.023	0.211	0.247	0.014
		Var(I)	0.234	0.028	0.024	0.183	0.293	-0.009
		Var(S)	0.014	0.002	0.018	0.011	0.017	0.004
		Cov(I, S)	-0.036	0.006	0.022	-0.049	-0.026	-0.005
		Var(e)	0.044	0.004	0.031	0.037	0.053	0.024
		df_y[5]	2.386	0.205	0.043	2.118	2.900	0.050
	Class 2	Intercept	6.196	0.047	0.020	6.103	6.287	0.054
		Slope	0.315	0.011	0.022	0.295	0.336	0.036
		Var(I)	1.326	0.084	0.017	1.167	1.497	0.020
		Var(S)	0.034	0.004	0.022	0.027	0.042	0.010
		Cov(I, S)	0.010	0.014	0.021	-0.018	0.037	-0.023
		Var(e)	0.372	0.020	0.033	0.336	0.412	-0.061
		df_y	3.200	0.195	0.040	2.850	3.600	-0.042
Probit Parameters	Class	φ_{10}[6]	-0.214	0.119	0.051	-0.438	0.018	-0.039
		φ_{11}	-0.223	0.077	0.051	-0.372	-0.076	0.026
	Grade 7	γ^*_{01}[7]	-0.711	0.532	0.066	-1.843	0.204	-0.255
		γ^*_{11}[8]	-0.132	0.216	0.058	-0.527	0.310	0.231
		γ_{x1}[9]	-0.154	0.108	0.046	-0.368	0.058	0.008
		γ_{Y1}[10]	-0.087	0.059	0.065	-0.190	0.038	0.251
	Grade 8	γ^*_{02}	-1.157	0.446	0.064	-2.097	-0.447	-0.373
		γ^*_{12}	0.046	0.217	0.055	-0.345	0.489	0.347
		γ_{x2}	0.113	0.114	0.046	-0.109	0.334	0.032
		γ_{Y2}	-0.108	0.045	0.062	-0.188	-0.021	0.330
	Grade 9	γ^*_{03}	-0.613	0.454	0.065	-1.519	0.163	-0.462
		γ^*_{13}	-0.057	0.181	0.056	-0.403	0.292	0.381
		γ_{x3}	-0.147	0.094	0.046	-0.332	0.038	0.045
		γ_{Y3}	-0.074	0.045	0.064	-0.155	0.022	0.459
	Grade 10	γ^*_{04}	-0.032	0.512	0.066	-0.861	0.985	-0.426
		γ^*_{14}	-0.324	0.204	0.059	-0.732	0.029	0.362
		γ_{x4}	0.059	0.101	0.047	-0.142	0.251	0.128
		γ_{Y4}	-0.166	0.050	0.065	-0.266	-0.084	0.378
	Grade 11	γ^*_{05}	-1.298	0.421	0.065	-2.130	-0.442	-0.192
		γ^*_{15}	0.341	0.176	0.055	0.015	0.708	0.159
		γ_{x5}	-0.087	0.091	0.045	-0.263	0.083	0.001
		γ_{Y5}	-0.019	0.040	0.064	-0.092	0.062	0.189

1 Ratio of MC error to standard deviation. A value around or less than 0.05 indicates that the corresponding estimate is accurate (Spiegelhalter, Thomas, Best, & Lunn, 2003).

2-3 The lower 2.5 percentile and upper 97.5 percentile.

4 Geweke test z value. An absolute value less than 1.96 indicates that the corresponding chain has passed the convergence test.

5 The degrees of freedom of the multivariate-t.

6 The probit coefficient of the class probability for class 1, defined in Eqn.(3).

7 The probit coefficient of the class membership 1 at Grade 7, defined in Eqn.(9).

8 The probit coefficient of the class membership 2 at Grade 7, defined in Eqn.(9).

9 The probit coefficient of the covariate at Grade 7, defined in Eqn.(9).

10 The probit coefficient of the potential output Y at Grade 7, defined in Eqn.(9).

use uninformative priors which carry little information for model parameters. A burn-in period is run first to ensure estimates are based on the Markov chains that have converged. For testing convergence, the history plot is examined and the Geweke's z statistic (Geweke, 1992) is checked for each parameter. The Geweke's z statistics for all the parameters are smaller than 1.96, which indicates converged Markov chains. To make sure all the parameters are estimated accurately, the next $50,000$ iterations are then saved for data analysis. The ratio of Monte Carlo error (MCerror) to standard deviation (S.D.) for each parameter is smaller than or close to 0.05, which indicates parameter estimates are accurate (Spiegelhalter, Thomas, Best, & Lunn, 2003). In step 3, model selection indices are used to compare the ten models. The indices are listed in Table 9. And in step 4, the results obtained from the final selected model are interpreted.

As suggested by Dhat.CAIC, Dhat.ssBIC, Dhat.BIC, and Dhat.AIC, without further substantive information, the TN-CXY model appears to be a good candidate for the best-fitting model. Table 10 provides the results of the TN-CXY REGMM model. It can be seen that (1) class 1 has a higher average initial level but a smaller average slope; (2) class 2 has larger variations for initial levels and slope; (3) the residual variance of class 2 is much larger than that of class 1; (4) in class 1 the initial level and the slope are significantly negatively correlated at the confidence level of 95%; (5) the missingness is not related to gender because none of the coefficients of gender are significant at the α level of 0.05; (6) at grade 11, adolescents in class 2 are more likely to miss tests than those in class 1 because the probit coefficient of class membership for grade 11 is significantly positive; and (7) at grades 8 and 10, students with higher potential scores are more likely to miss tests than the students having lower scores because the probit coefficients of the potential outcomes y at the two grades are significantly negative.

6 Conclusions, Discussion and Future Research

Based on the results from the five simulation studies, one can conclude that (1) almost all of the model selection indices, except for the rough DIC (RDIC), can correctly choose the true model with high certainty; (2) if the number of classes is correctly identified, then the Dbar-based indices perform better than the Dhat-based indices; if candidate models have different numbers of classes, then the Dhat-based indices might be used to select the best fit model; (3) across 5 studies, CAIC and BIC provide higher probabilities than those ssBIC, AIC, or DIC does. The results will help inform the selection of growth models by researchers seeking to provide people with accurate estimates of growth across a variety of possible contexts. The real data analysis demonstrated the application of the indices to typical longitudinal growth studies such as educational, psychological, and social research.

This study can be extended in many ways. For example, different versions of the likelihood function or more model selection indices can be studied and compared by using more practical statistical models. (1) As we stated in the section of Introduction, there are at least three challenges in proposing

new selection indices. The third challenge is about the likelihood function $l(y|\hat{\theta})$. When latent variables involved, the likelihood can be an observed-data likelihood, a complete-data likelihood, or a conditional likelihood (Celeux et al., 2006). In this study, we use a conditional joint loglikelihood, but in the future, the other versions of likelihood functions can be investigated. (2) Another future research of this study is to propose other model selection indices, such as Bayes factors. (3) This study focuses on latent growth models only. In the future, the performance of these selection indices can be studied by using other statistical models, such as survival models.

Acknowledgment

The study was supported by a grant from the Institute of Education Sciences (R305D210023).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *1919 (6)*, 716-723. doi: https://doi.org/10.1007/978-1-4612-1694-0_16

Anderson, T. W., & Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *The Annals of Mathematical Statistics*, *33*, 420-431. doi: https://doi.org/10.1214/aoms/1177704568

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis: Kendall's library of statistics* (2nd ed., Vol. 7). New York, NY: Edward Arnold.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370. doi: https://doi.org/10.1007/bf02294361

Bureau of Labor Statistics, U.S. Department of Labor. (1997). *National longitudinal survey of youth 1997 cohort, 1997-2003 (rounds 1-7).* [computer file]. Produced by the National Opinion Research Center, the University of Chicago and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2005. Retrieved from `http://www.bls.gov/nls/nlsy97.htm`

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174. doi: https://doi.org/10.2307/2685208

Celeux, G., Forbes, F., Robert, C., & Titterington, D. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *4*, 651-674. doi: https://doi.org/10.1214/06-ba122

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, B*, *62*, 355-366. doi: https://doi.org/10.1111/1467-9868.00236

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (p. 169-193). Oxford, UK: Clarendon Press.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Drawing inferences from self-selected samples. In H. Wainer (Ed.), (p. 115-142). New York: Springer Verlag.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125-134. doi: https://doi.org/10.2307/2290705

Little, R. J. A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121. doi: https://doi.org/10.1080/01621459.1995.10476615

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, N.Y.: Wiley-Interscience. doi: https://doi.org/10.1002/9781119013563

Lu, Z., & Zhang, Z. (2014). Robust growth mixture models with non-ignorable missingness data: Models, estimation, selection, and application. manuscript submitted for publication. *Computational Statistics and Data Analysis*, *71*, 220–240. doi: https://doi.org/10.1016/j.csda.2013.07.036

Lu, Z., & Zhang, Z. (2021). Bayesian approach to non-ignorable missingness in latent growth models. *Journal of Behavioral Data Science*, *1*(2), 1–30. doi: https://doi.org/10.35566/jbds/v1n2/p1

Lu, Z., Zhang, Z., & Cohen, A. (2013). New developments in quantitative psychology. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), (Vol. 66, p. 275-304). Springer New York.

Lu, Z., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent-class-dependent missing data. *Multivariate Behavioral Research*, *46*, 567-597. doi: https://doi.org/10.1080/00273171.2011.589261

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. John Wiley & Sons, Inc. doi: https://doi.org/10.1002/0470010940

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons. doi: https://doi.org/10.1002/0471721182

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*(2), 463-469. doi: https://doi.org/10.1111/j.0006-341x.1999.00463.x

Oldmeadow, C., & Keith, J. M. (2011). Model selection in Bayesian segmentation of multiple DNA alignments. *Bioinformatics*, *27*, 604-610. doi: https://doi.org/10.1093/bioinformatics/btq716

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6 (2)*, 461-464. doi: https://doi.org/10.1214/aos/1176344136

Sclove, L. S. (1987). Application of mode-selection criteria to some problems in multivariate analysis. *Psychometrics*, *52*, 333-343. doi:

https://doi.org/10.1007/bf02294360

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. v. d. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583-639. doi: https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS manual Version 1.4.

(MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK, http://www.mrc-bsu.cam.ac.uk/bugs)

Yuan, K.-H., & Lu, Z. (2008). SEM with missing data and unknown population using two-stage ML: theory and its application. *Multivariate Behavioral Research*, *43*, 621-652. doi: https://doi.org/10.1080/00273170802490699

Zhang, Z., Lai, K., Lu, Z., & Tong, X. (2013). Bayesian inference and application of robust growth curve models using student's t distribution. *Structural Equation Modeling*, *20(1)*, 47-78. doi: https://doi.org/10.1080/10705511.2013.742382

Journal of Behavioral Data Science, 2022, *2* (1), 59–74.
DOI: https://doi.org/10.35566/jbds/v2n1/p3

Does Minority Case Sampling Improve Performance with Imbalanced Outcomes in Psychological Research?

Ross Jacobucci*[0000−0001−7818−7424] and Xiaobei Li[0000−0002−8266−9490]

University of Notre Dame, Notre Dame, IN 46530, USA
rjacobuc@nd.edu, xli29@nd.edu

Abstract. In psychological research, class imbalance in binary outcome variables is a common occurrence, particularly in clinical variables (e.g., suicide outcomes). Class imbalance can present a number of difficulties for inference and prediction, prompting the development of a number of strategies that perform data augmentation through random sampling from just the positive cases, or from both the positive and negative cases. Through evaluation in benchmark datasets from computer science, these methods have shown marked improvements in predictive performance when the outcome is imbalanced. However, questions remain regarding generalizability to psychological data. To study this, we implemented a simulation study that tests a number of popular sampling strategies implemented in easy-to-use software, as well as in an empirical example focusing on the prediction of suicidal thoughts. In general, we found that while one sampling strategy demonstrated far worse performance even in comparison to no sampling, the other sampling methods performed similarly, evidencing slight improvements over no sampling. Further, we evaluated the sampling strategies across different forms of cross-validation, model fit metrics, and machine learning algorithms.

Keywords: Imbalanced data · Sampling strategies · Machine learning

1 Introduction

In psychological research, class imbalance in binary outcome variables (also referred to as skew or rare events), most often occurs due to the underlying population of interest having small proportions of individuals with positive cases (minority), such as in the case of study designs that are assessing the prevalence of suicidal attempts in the general population. While class imbalance can be dealt with through changes in study design, such as sampling among individuals with a history of mental illness to increase the probability of observations having a history of suicide attempts, this can fundamentally alter the alignment

between the population of interest and the sample of which the data is collected from.

In the presence of class imbalance, failure to utilize appropriate strategies has a number of consequences. For instance, even when explanation is the primary aim, using logistic regression with skewed outcomes can result in underestimated probabilities for the positive class (King & Zeng, 2001). Additionally, one of the general strategies is to perform data augmentation through random sampling from just the positive cases, or from both the positive and negative cases. Kovács (2019) found that in general, any form of sampling improves upon the performance modeling the original dataset.

However, even in areas of psychological research where imbalanced outcomes are extremely common, such as suicide, the use of sampling strategies is still extremely rare (e.g., a recent tutorial on evaluating classification in suicide research does not mention sampling strategies; Mitchell, Cero, Littlefield, & Brown, 2021). While part of this lack of translation across disciplines may be due to relatively siloed research, it also may partially be attributed to a lack of generalizability in the findings. While a large number of studies have evaluated the relative performance of methods designed to overcome class imbalance, the vast majority of research focuses on evaluation in benchmark datasets with characteristics unique to that field of study (primarily computer science), which limits the generalizability of these findings to areas with different types of data including psychology. This is similar to the hype and promise of machine learning being somewhat diluted by limitations to the data commonly found in psychological research (see Jacobucci & Grimm, 2020).

Thus, the goal of this study is to answer the question: Do minority case sampling approaches improve prediction with imbalanced outcomes in datasets with psychological variables? To accomplish this, we evaluated a number of sampling strategies commonly used for imbalanced outcomes in simulated data that is more in line with characteristics commonly found in psychological research. We followed this by applying those strategies to the prediction of suicidal ideation in a large public dataset. Additionally, we specifically focus on strategies for overcoming class imbalance that are already implemented in easy-to-use software. Our focus is on strategies that operate at the data level, as opposed to the model estimation phase. While we test two different algorithms, logistic regression and random forests, we put our focus on methods that do not involve the use of misclassification costs for a number of reasons. The first is that sampling methods are easy to implement in easy-to-use software that pairs with many ML algorithms, meaning researchers won't face limitations with which algorithms can be compared. Further, sampling based methods have been studied more (e.g., García, Sánchez, Marqués, Florencia, & Rivera, 2020) thus often show up more in recommendations. And finally, assigning costs does not overcome potential issues of few to no positive cases being represented when the sample size is small and k-fold cross-validation (CV) is used.

1.1 Sampling Methods

While a number of strategies have been proposed for supervised learning with imbalance data, possibly the two simplest are random over-sampling (OVER) and random under-sampling (UNDER). While OVER random samples from the minority case to produce an equal distribution of positive and negative cases, UNDER randomly removes majority cases to produce an equal distribution. As an example, of an original dataset with 10 positive cases and 100 negative cases, OVER would produce a new dataset with 100 positive and negative cases each, while UNDER would create a dataset with 10 positive and negative cases each. Both methods have well understood drawbacks: while UNDER discards potentially useful data, OVER increases the probability of overfitting (McCarthy, Zabar, & Weiss, 2005).

1.2 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) is an oversampling technique that creates artificial minority class cases by using the k-nearest neighbors for a given minority case instead of oversampling randomly as in OVER. More specifically, for a specific minority case i, instead of just creating duplicates of that case, the SMOTE algorithm finds k similar minority cases to case i, and generates synthetic cases that take on a value for the predictor variables that represent a blend of the k-nearest neighbors. Thus, the newly created synthetic minority cases contain similar, not identical, predictor values to the k-nearest neighbors. Finally, the number of synthetic cases created for each minority case is a tuning parameter typically referred to as $N_{percent}$.

1.3 Random Over-Sampling Examples (ROSE)

While SMOTE only generates synthetic samples from the positive cases, ROSE (Menardi & Torelli, 2012) uses the smoothed bootstrap to generate both negative and positive samples to create a new dataset that is more balanced. In the ROSE procedure, an observation is first drawn with a 50% chance of belonging to each class. Given this observation, a new sample is then generated in its neighborhood (according to the predictor values), with the neighborhood chosen according to a kernel density estimate (for further detail, see Menardi & Torelli, 2012). The user of ROSE is given discretion as to how much under-sample the negative cases, and to what degree over-sample the positive cases.

Demonstration To demonstrate how SMOTE and ROSE randomly over sample minority cases, we simulated 50 cases according to a linear logistic model with two predictors (regression coefficients of 0.2 and 0.4) and an intercept of -3, which resulted in 47 negative cases and 3 positive cases. This was followed by applying each method with the `DMwR` package (Torgo, 2010) and `ROSE` package (Lunardon, Menardi, & Torelli, 2014). The resulting datasets are displayed in Figure 1.

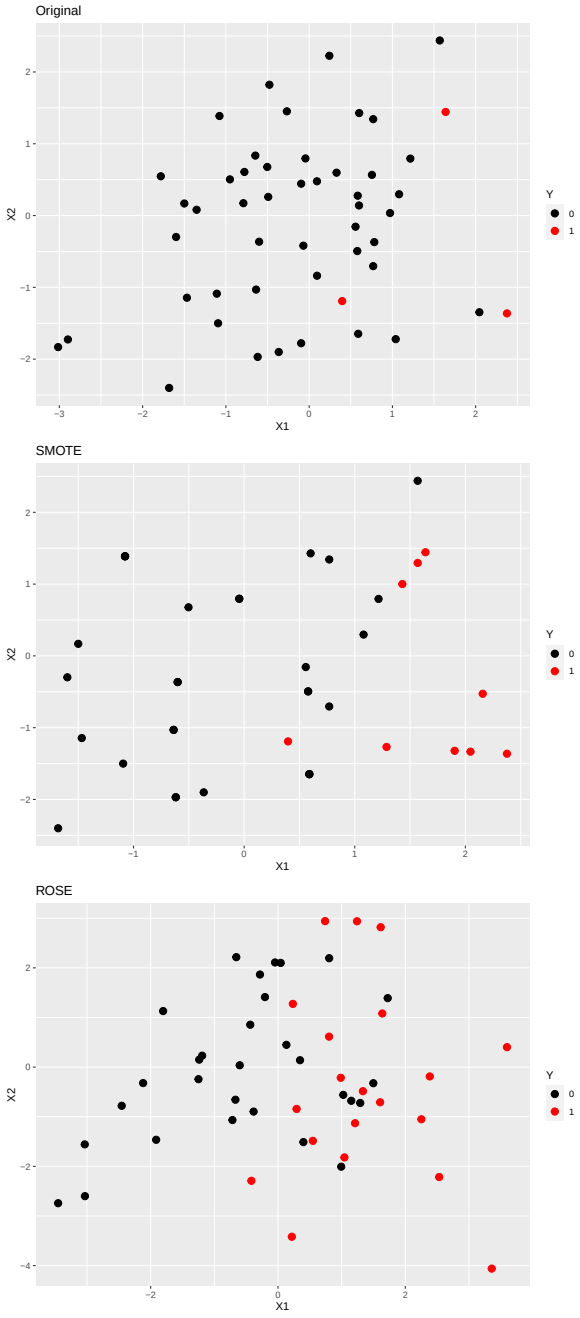


Figure 1: The top plot is the original dataset, the middle figure is after applying SMOTE, while the bottom figure is after applying ROSE.

We can see in Figure 1 that both methods oversample the positive cases, generating new data points based on the original positive cases. Additionally, the two methods handle the negative cases in different ways: SMOTE uses a hyperparameter (we selected 500%) to select the percentage of negative cases that are selected relative to the number of positive cases generated; ROSE requires setting a hyperparameter to determine the percentage of positive cases that end up in the new dataset (defaults to 0.5), with the negative cases making up the remainder to have the sample size equal to the original sample.

1.4 Comparison of Sampling Procedures

SMOTE has been widely applied across an assortment of research domains. Evaluating its performance in applied data is complicated by a number of factors, but one primary concern is on the differences in which classification performance metrics are reported. For instance, Nnamoko and Korkontzelos (2020) used data with diabetes diagnosis as the outcome, but mainly focused on improvements due to SMOTE in accuracy. However, the calculation of accuracy is based on the class distribution, which is discrepant across the various ways that Nnamoko and Korkontzelos evaluated the use of SMOTE. Most studies have primarily focused on evaluating the area under the receiver operator characteristic (ROC) curve (AUC; Hanley & McNeil, 1982), or the area under the precision-recall curve (AUPRC; He & Garcia, 2009). There exist multiple additional methods that incorporate similar types of information (e.g., see Saito & Rehmsmeier, 2015), however, these two have received the most coverage. Whereas the AUC encompasses the contrast between sensitivity and specificity, thus information regarding both classes, the AUPRC contrasts recall with precision, thus only encodes information regarding positive cases. While the AUC is more commonly used in practice, there are concerns regarding the AUC being misleading in the presence of imbalance (Lobo, Jiménez-Valverde, & Real, 2008), with findings that the AUPRC is more informative when classes are imbalanced (Saito & Rehmsmeier, 2015). Therefore, we will focus on both the AUC and AUPRC, but give preference to the AUPRC.

An additional complication in evaluating methods for handling imbalance is the research domain of concern. Many studies that evaluate methods for handling imbalance use benchmark datasets from that research area. For instance, a recent study by Shin et al. (2021) examined the bloom of cyanobacteria in rivers in South Korea. With this data, the researchers found differences in performance among various classifiers (e.g., ensembles outperformed single models), but only marginal performance gains in the application of SMOTE. Notably, they did not evaluate the AUPRC. In a different area of application, Zhu, Baesens, and vanden Broucke (2017) examined class imbalance strategies in the area of customer churn prediction, evaluating performance in terms of the AUC, and comparing ensemble methods paired with various sampling strategies and cost-sensitive learning. Again, ensemble methods outperformed simpler algorithms, however, there did not seem to be a benefit to more complex sampling strategies above and beyond over or under-sampling. Finally, Demir and Şahin (2022) examined the

impact of classification algorithms and oversampling methods for soil liquefaction evaluation, finding that SMOTE outperformed both OVER and ROSE.

1.5 General Findings

Among relatively simple strategies for handling class imbalance, over-sampling is typically preferred over under-sampling (e.g., Batista, Prati, & Monard, 2004; Buda, Maki, & Mazurowski, 2018). In simulated and benchmark datasets, García et al. (2020) compared under-sampling, over-sampling, and a hybrid of both to just the use of the original dataset, confirming prior research that over-sampling outperforms under-sampling.

A further complication in this is that originally proposed over-sampling methods can be subject to different interpretation, resulting in varying implementations. To address this, Bajer, Zonc, Dudjak, and Martinovic (2019) tested four possible variants on the original SMOTE implementation, along with more recently proposed generalizations of SMOTE. On a number of benchmark datasets, they found that all of the variants outperformed random over-sampling and no sampling, with the highest performance attributed to the recently proposed Weighted-SMOTE (Prusty, Jayanthi, & Velusamy, 2017).

A recent study attempting to provide benchmark performance metrics for a host of recently proposed advancements on a large number of benchmark datasets, evaluated with multiple ML algorithms with repeated k-fold cross-validation (Kovács, 2019). They found that the biggest improvements were attributed to the use of any reliable over-sampling method over no sampling, with much smaller improvements due to the use of the best performing methods over standard SMOTE oversampling. However, importantly for our purposes, this study did not test random over-sampling, and like most other studies, used a large number of benchmark datasets.

One key piece in applying over-sampling is to ensure that augmented datasets are not created prior to splitting the dataset up into training and tests sets, as this can lead to overly-optimistic performance due to data leakage (e.g., Vandewiele et al., 2021). This can be attributed to copies, either exact or very similar, of original cases being included in both the training and tests sets. In `R` package `caret` (Kuhn, 2008), the resampling is conducted inside of cross-validation or bootstrap sampling. As an example, in 5-fold CV, each partition that is created with 4/5ths of the sample is then subject to over-sampling, the model is trained, then tested on the 1/5th sample that was not subject to sampling. However, of note, if one has a true test set that is only used for assessing the final model's performance, sampling should not be conducted in this sample, as it is used only to test a previously trained model.

Finally, much less research has focused on the interaction between the use of sampling and the actual sample size of the dataset. Studying this interaction is further complicated by the form of resampling used to evaluate prediction performance, as the most commonly used form, k-fold CV, has been shown to produce highly biased results in small samples (Vabalas, Gowen, Poliakoff, & Casson, 2019). However, the presence of a binary outcome further complicates

defining what sample size is given that assessing the number of positive cases is more informative than the overall number of cases (cf. Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996).

2 Study 1

2.1 Methods

We specifically chose this study design as we believe that it mimics the structure found in the majority of clinical data that primarily includes self-report data. While we assess the influence of nonlinear relationships, we primarily simulate the data according to a linear model, as this is most in line with the results that utilize machine learning algorithms in clinical self-report data: if machine learning outperforms linear models, the improvement in performance is most often negligible (Christodoulou et al., 2019; Jacobucci, Littlefield, Millner, Kleiman, & Steinley, 2021). For the simulation setup, we started by simulating standard normally distributed data with a sample size of 50,000. The cases not selected to train and test the methods were kept in order to produce performance metrics that serve as ground truth.

In predictive tasks with class outcomes, there are often two layers of assessment. The first step in prediction-oriented tasks is often assessing the correspondence between the predicted probabilities and actual class labels, while further performance assessment can be taken in translating the predicted probabilities to predicted class labels to classify individuals. Given that much of psychological research is only focused with the first step, our aim is only assessing prediction performance. We interpret performance with respect to the AUC and AUPRC. While the AUPRC is more informative at higher degrees of imbalance, the AUC is much less likely to evidence floor effects, thus improving our ability to characterize its distribution. When making specific comparisons in performance across methods, we used an ANOVA with Tukey's HSD posthoc tests.

With this setup, we varied a number of conditions across 200 repeats: To train and test the methods we tested sample sizes of 300, 1000, and 10,000. With these sample sizes, we simulated data following a logit link, while varying the intercept (b_0) to control the level of class imbalance. We specifically tested values of -4 ($\approx$ 0.02 positive), -3 ($\approx$ 0.05 positive), -2 ($\approx$ 0.12 positive), -1 ($\approx$ 0.27 positive), and 0 ($\approx$ balanced case). We tested the inclusion of 30 and 70 predictors, with 10% of the predictors having standardized coefficients of 0.2, 10% having 0.1, 10% having 0.05, and the rest 70% having coefficients of 0. Additionally, we added two standard normal predictors with unit weighted cosine and sine relationships with Y, and a *tanh* interaction between these variables with coefficients of 0.1. Although the exact functional form of these nonlinear relationships is unlikely to occur in psychological data, our focus is less on identifying the true relationships and more on determining whether nonlinearity interacted with imbalance to bias our model performance or algorithm selection. Following the logistic model, we also tested residual variances of 0.82 and 0.3. Once this

normally distributed version of Y was created, it was transformed into a probability according to a logit link to a probability, followed by using a binomial distribution to generate values of 0 and 1. Finally, we included Bayesian logistic regression (Bayesian as this resulted in fewer convergence issues when the class imbalance was large and sample size small) and random forests (Breiman, 2001). Our goal in this comparison was to test the potential of underfitting and overfitting, particularly given prior findings regarding overfitting with ROS. Our specific point of comparison is in assessing the performance of random forests with sampling to determine whether inflated AUC or AUPRC values are found.

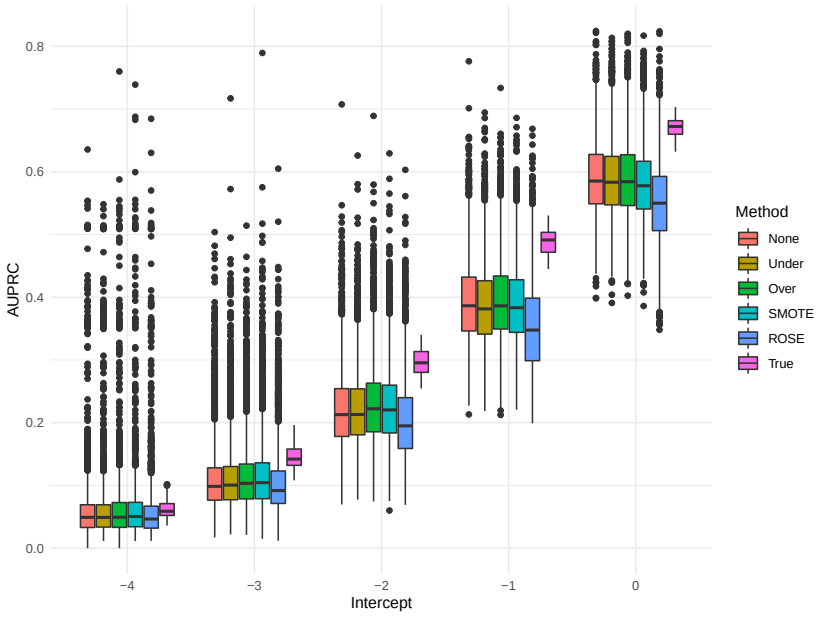
Outside of the simulated intercept values, our main point of evaluation concerned the resampling and sampling methods. For resampling, we tested the validation set approach or 10-fold CV. The validation set approach applied the sampling method on a training set that contained 70% of N, followed by testing on a holdout set containing 30% of N that was not used to train the model. The 10-fold CV approach used the sampling approach on each training set for each 10 iterations. Note that for both approaches the sampling method was used after splitting the sample and was not applied to the holdout set. Finally, our goal in assessing sampling methods was to test methods that are easy to apply in commonly used software. Given this, we focused on the methods available in the `caret` package (Kuhn, 2008) in `R`. This included no sampling, UNDER, OVER, SMOTE, and ROSE. SMOTE is implemented in the `DMwR` package (Torgo, 2010), while ROSE is implemented in the `ROSE` package (Lunardon et al., 2014). We used the software defaults for both SMOTE and ROSE.

2.2 Results

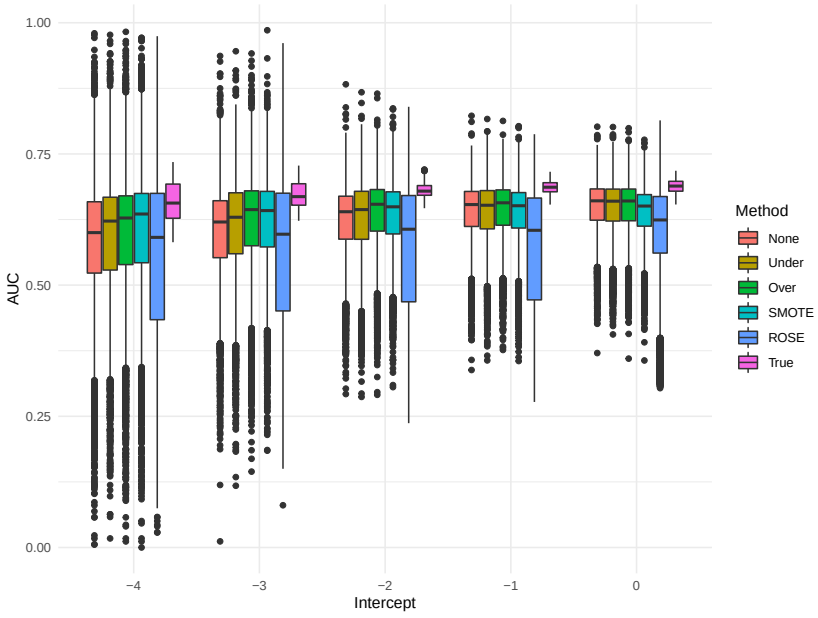
Of the 200 replications across conditions, errors occurred in estimating models in a subset of conditions, namely due to the condition with a sample size of 300 and intercept of -4 (11% errors). All other conditions had less than 2% errors. Given the breadth of results, we chose only to present a select subset of findings that highlight key points.

Intercept and variability Possibly the largest influence of class imbalance is on the degree of variability to the AUC and AUPRC. This can be clearly seen in Figure 2. With an intercept of -4, all of the sampling methods evidence a number of outliers that are strongly positively biased. However, this represented a quite small number of results, as the 95th percentile at an intercept of -4 was 0.13 for OVER and SMOTE.

Additionally, we can see the median AUC and AUPRC values for the sampling methods get closer to the True performance as the imbalance becomes less. This is further influence by sample size and can be attributed to a lack of information when there are fewer positive cases, leading to further degrees of underfitting by default. As an example, when the intercept is -4, there is a difference in AUC means of 0.06 between OVER and the True performance (0.596 vs. 0.660), while for an intercept of 0 it is 0.04 (0.648 vs. 0.688).



(a) AUPRC



(b) AUC

Figure 2: AUPRC and AUC values across sampling methods and simulated intercept.

Lastly, while the AUC is commonly labeled as biased in the presence of imbalanced data, it demonstrated similar performance across values of b_0 as the AUPRC. Performance improves as the class distribution becomes more equal; this is most likely attributable to the increasing numbers of positive cases. In fact, when the intercept was -4, the AUC was on average 0.09 points higher when the sample size was 10,000 as opposed to 300, while the discrepancy fell to 0.06 when the intercept was 0.

We see similar effects with sample size in Figure 3, as one would expect. Larger sample size resulted in less variability, which further reduced the propensity to over- or under-estimated performance. For the AUC [1], the standard deviations were 0.10 for 300, 0.07 for 1,000, and 0.07 for 10,000. While a large number of AUC values for each of the sampling methods were greater than the True values in the smaller sample, the median values became closer to the True median scores in the larger sample sizes. This highlights improvements in performance and stability with greater N, and a worrisome level of biased outliers at small sample sizes.

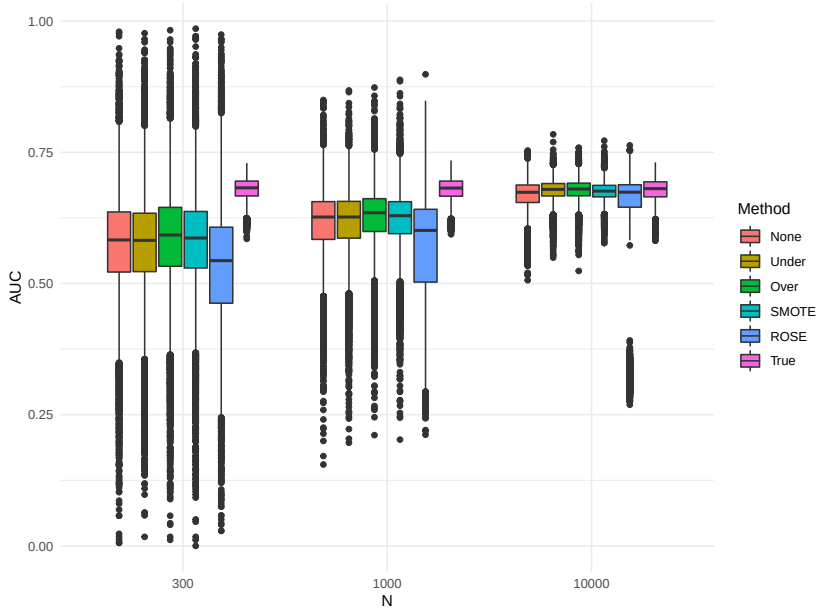


Figure 3: Sample size and AUC values. Note that we don't also depict the AUPRC as the behavior is the same, but the variance is overly wide due to averaging over the intercept values.

[1] We don't report the standard deviations for the AUPRC as there were floor effects.

k-fold CV reduces variability Figure 4 displays the AUC values across both k-fold CV and validation set strategies, as well as using random forests and logistic regression. The first thing to note is the differences in variability across resampling methods, with k-fold CV having lower variance, particularly at a sample size of 300. This is in line with general recommendations to only use the validation set strategy in the presence of large sample sizes (i.e., James, Witten, Hastie, & Tibshirani, 2013). Secondly, there do not seem to be mean differences across resampling methods, and only slight improvements due to random forests (as expected). Finally, we can see a strange interaction between the use of ROSE with k-fold CV, resulting in markedly worse performance.

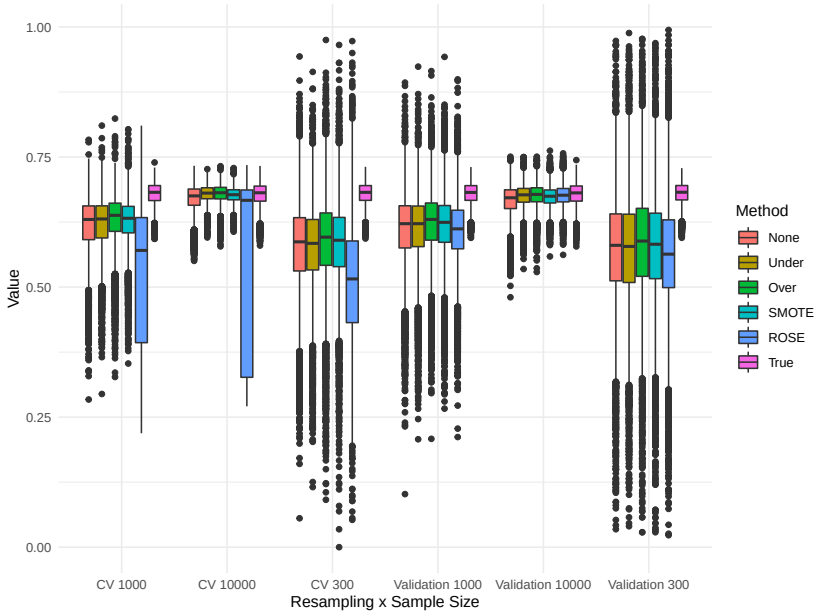


Figure 4: Sample size and AUC values. Note that the AUPRC is not depicited as the behavior is the same, but the variance is overly wide due to averaging over the intercept values.

Summary There was no form of sampling that resulted in universal best performance, but some general trends emerged. In line with prior research, some form of sampling generally outperformed no sampling. For both the linear and nonlinear simulated data, OVER had higher AUC values by 0.01 across conditions, which held even when imbalance was at its highest ($b_0 = -4$) and a sample size of 300. Further, OVER and SMOTE consistently performed the best, with no statistical difference in their results averaged across conditions. For instance,

when the intercept was -4, OVER had a median AUPRC of 0.057, SMOTE of 0.058, while no sampling was 0.052. It is important to note however that there was more variability within each type of sampling than between methods. Of note, there was not a notable distinction in the amount of performance variance across sampling methods: unfortunately, all sampling methods evidenced a large degree of variability when the sample size was small (Figure 3) or there was high class imbalance (Figure 2). Finally, ROSE performed the worst among the sampling methods, which was primarily due to problems in integrating ROSE with logistic regression and k-fold CV as seen in Figure 4.

A surprising finding was that there were very little differences in the variability between the use of logistic regression and random forests. Random forests performed better than logistic regression, as expected given the two nonlinear effects, but importantly random forests did not have a greater propensity to overfit than logistic across the conditions and sampling methods.

3 Study 2

Data for Study 2 comes from the National Survey on Drug Use and Health from 2014 (NSDUH; Abuse & Administration., n.d.). This survey focused on assessing the use of illicit drugs, alcohol, and tobacco among U.S. civilians 12 years or older. For the purpose of our analysis, we focused on questions that assessed mental health issues. With a sample size of 55,271 and 3,148 variables, the dataset was pared down from the original dataset to just include thirty-nine predictors with the aim of predicting suicidal ideation (last 12 months; SUICTHINK). Predictors included symptoms of depression and other mental health disorders, the impact of these symptoms on daily functioning, and four demographic variables (gender, ethnicity, relationship status, age; dummy coded). The dataset can be freely downloaded from https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2014-nsduh-2014-ds0001-nid16876.

For the analysis, we used Bayesian logistic regression and random forests, while testing all of the above forms of handling imbalance. Secondly, we detail both the AUPRC and AUC given that the outcome variable had only 3.7% positive cases. Additionally, we separate the results by whether the sampling method for handling imbalance paired with the validation set approach or 10-fold CV. Finally, we do not report the results using ROSE given its poor performance in the simulation.

3.1 Results

As seen in Table 1, almost uniformly, the AUC and AUPRC values were higher when using 10-fold CV as opposed to the validation set approach, highlighting again that when comparing results across algorithms the same resampling strategy should be used. In assessing the AUC, OVER sampling performed slightly better than no sampling, while the opposite was true for the AUPRC. In fact, no sampling had the highest AUPRC values. In digging deeper to the simulation

results, at a sample size of 10,000, there were no statistical differences across the sampling methods for the AUPRC, with a similar lack of distinction even at smaller sample sizes. This empirical example further highlights that at large sample sizes, the use of sampling methods matters less, particularly when using the AUPRC.

Table 1: Results from the Empirical Analysis.

	AUC				AUPRC			
	None	UNDER	OVER	SMOTE	None	UNDER	OVER	SMOTE
	Logistic Regression							
Validation	0.801	0.805	0.805	0.772	0.302	0.289	0.301	0.224
10-Fold	0.809	0.806	0.810	0.807	0.321	0.299	0.317	0.308
	Random Forest							
Validation	0.761	0.772	0.765	0.764	0.281	0.268	0.245	0.267
10-Fold	0.768	0.781	0.771	0.774	0.304	0.267	0.260	0.283

4 Conclusion

This paper addressed a number of decision points that psychological researchers face when analyzing outcomes that exhibit imbalance. These decision points are particularly relevant when applying machine learning, as the importance of cross-validation and the accurate testing of hyperparameters become increasingly important. With this, there were a number of key takeaways:

- k-fold CV should be preferred to the validation set approach when using sampling methods to address class imbalance.
- The AUC did not demonstrate a bias in the presence of imbalance when using as an overall metric of fit (as opposed to examining the ROC curve).
- While OVER, UNDER, and SMOTE sampling approaches demonstrated improvements of no sampling, these improvements were extremely small.
- The use of sampling did not increase the propensity to overfit, even when paired with random forests.
- The ROSE method should not be used.
- Simple models such as logistic regression may outperform complex machine learning algorithms in predicting psychological phenomena (i.e., Jacobucci & Grimm, 2020).

Additionally, although the use of sampling can improve mean/median estimates of performance in the presence of imbalance, there were not meaningful reductions in variability to the performance estimates. This finding would not have been identified by following the standard use of benchmark datasets, and is only possible through the use of simulation.

While this study was able to answer a number of questions, there are some important limitations. The first is that the data was simulated in a relatively simple way, following a logistic link with standard normal variables. Therefore, there remains uncertainty as to how the methods perform with datasets that exhibit levels of complexity falling in between our simulated data approach and the benchmark datasets commonly used to test the sampling approaches. A second limitation is that we only tested the sampling approaches that are easily applied using the `caret` package in `R`, while prior research has found performance improvements in a number of more recently developed approaches, particularly generalizations of SMOTE. While `R` users can write their own functions implementing the additional varieties of SMOTE to be paired with caret, this is unlikely to occur in the majority of psychological applications.

References

Abuse, S., & Administration., M. H. S. (n.d.). *National survey on drug use and health.* Retrieved from `http://www.samhsa.gov/data/population-data-nsduh/reports(accessedJuly28,2015).`

Bajer, D., Zonc, B., Dudjak, M., & Martinovic, G. (2019). Performance analysis of smote-based oversampling techniques when dealing with data imbalance. *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 265-271. doi: https://doi.org/10.1109/IWSSIP.2019.8787306

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, *6*(1), 20–29. doi: https://doi.org/10.1145/1007730.1007735

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. doi: https://doi.org/10.1023/A:1010933404324

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249-259. doi: https://doi.org/10.1016/j.neunet.2018.07.011

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, *16*(1), 321–357. doi: https://doi.org/10.1613/jair.953

Christodoulou, E., Ma, J., Collins, G., Steyerberg, E., Verbakel, J., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*. doi: https://doi.org/10.1016/j.jclinepi.2019.02.004

Demir, S., & Şahin, E. K. (2022). Evaluation of oversampling methods (over, smote, and rose) in classifying soil liquefaction dataset based on svm, rf, and naïve bayes. *Avrupa Bilim ve Teknoloji Dergisi*(34), 142 - 147. doi: https://doi.org/10.31590/ejosat.1077867

García, V., Sánchez, J., Marqués, A., Florencia, R., & Rivera, G. (2020). Understanding the apparent superiority of over-sampling through an analysis

of local information for class-imbalanced data. *Expert systems with applications*, *158*, 113026. doi: https://doi.org/10.1016/j.eswa.2019.113026

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29-36. doi: https://doi.org/10.1148/radiology.143.1.7063747

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. doi: https://doi.org/10.1109/TKDE.2008.239

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809-816. (PMID: 32348703) doi: https://doi.org/10.1177/1745691620902467

Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science*, *9*(1), 129-134. doi: https://doi.org/10.1177/2167702620954216

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (Vol. 103). New York: Springer. doi: https://doi.org/10.1007/978-1-4614-7138-7

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163. doi: https://doi.org/10.1093/oxfordjournals.pan.a004868

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied soft computing*, *83*, 105662. doi: https://doi.org/10.1016/j.asoc.2019.105662

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, *28*(5). doi: https://doi.org/10.18637/jss.v028.i05

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and biogeography*, *17*(2), 145–151. doi: https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lunardon, N., Menardi, G., & Torelli, N. (2014). Rose: a package for binary imbalanced learning. *The R journal*, *6*(1), 79. doi: https://doi.org/10.32614/RJ-2014-008

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? In *Conference on knowledge discovery in data: Proceedings of the 1st international workshop on utility-based data mining; 21-21 aug. 2005* (pp. 69–77). ACM. doi: https://doi.org/10.1145/1089827.1089836

Menardi, G., & Torelli, N. (2012). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, *28*(1), 92–122. doi: https://doi.org/10.1007/s10618-012-0295-5

Mitchell, S. M., Cero, I., Littlefield, A. K., & Brown, S. L. (2021). Using categorical data analyses in suicide research: Considering clinical utility

and practicality. *Suicide & life-threatening behavior*, *51*(1), 76–87. doi: https://doi.org/10.1111/sltb.12670

Nnamoko, N., & Korkontzelos, I. (2020). Efficient treatment of outliers and class imbalance for diabetes prediction. *Artificial intelligence in medicine*, *104*, 101815–101815. doi: https://doi.org/10.1016/j.artmed.2020.101815

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, *49*(12), 1373–1379. doi: https://doi.org/10.1016/s0895-4356(96)00236-3

Prusty, M. R., Jayanthi, T., & Velusamy, K. (2017). Weighted-smote: A modification to smote for event classification in sodium cooled fast reactors. *Progress in nuclear energy (New series)*, *100*, 355–364. doi: https://doi.org/10.1016/j.pnucene.2017.07.015

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432–e0118432. doi: https://doi.org/10.1371/journal.pone.0118432

Shin, J., Yoon, S., Kim, Y., Kim, T., Go, B., & Cha, Y. (2021). Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological informatics*, *61*, 101202. doi: https://doi.org/10.1016/j.ecoinf.2020.101202

Torgo, L. (2010). *Data mining with r, learning with case studies*. Chapman and Hall/CRC. doi: https://doi.org/10.1201/9780429292859

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365–e0224365. doi: https://doi.org/10.1371/journal.pone.0224365

Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., ... Demeester, T. (2021). Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, *111*, 101987. doi: https://doi.org/https://doi.org/10.1016/j.artmed.2020.101987

Zhu, B., Baesens, B., & vanden Broucke, S. K. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information sciences*, *408*, 84–99. doi: https://doi.org/10.1016/j.ins.2017.04.015

Journal of Behavioral Data Science, 2022, *2* (1), 75–105.
DOI: https://doi.org/10.35566/jbds/v2n1/p5

The Impact of Sample Size on Exchangeability in the Bayesian Synthesis Approach to Data Fusion*

Katerina M. Marcoulides*[1], Jia Quan[2], and Eric Wright[2]

[1] Department of Psychology, University of Minnesota
kmarcoul@umn.edu
[2] University of Florida

Abstract. Data fusion approaches have been adopted to facilitate more complex analyses and produce more accurate results. Bayesian Synthesis is a relatively new approach to data fusion where results from the analysis of one dataset are used as prior information for the analysis of the next dataset. Datasets of interest are sequentially analyzed until a final posterior distribution is created, incorporating information from all candidate datasets, rather than simply combining the datasets into one large dataset and analyzing them simultaneously. One concern with this approach lies in the sequence of datasets being fused. This study examines whether the order of datasets matters when the datasets being fused each have substantially different sample sizes. The performance of Bayesian Synthesis with varied sample sizes is evaluated by examining results from simulated data with known population values under a variety of conditions. Results suggest that the order in which the dataset are fused can have a significant impact on the obtained estimates.

Keywords: Bayesian synthesis · Data fusion · Exchangeability

1 Introduction

Researchers in psychology and in the social and behavioral sciences more broadly, have recently expressed concerns about a "replication crisis" (Maxwell, Lau, & Howard, 2015). These concerns have driven researchers to explore and develop new strategies for analyzing data across multiple studies and summarizing results. In an effort to combat the replication crisis, open-source data repositories have grown substantially (Bhattacharya & Saha, 2015), and this greater access

* Early versions of this simulation study were presented at the International Meeting of the Psychometric Society and the Annual Meeting of the Florida Educational Research Association. This paper extends the simulation conditions in the first author's dissertation submitted in partial fulfillment of the doctoral degree at Arizona State University, Tempe.

to data both enables and requires the development of new strategies for exploring and analyzing this large amount of publicly available data. Data fusion is one method that has been shown to enable more complex and more appropriate models to be fit to fused (i.e., combined) datasets than just a single dataset (Curran & Hussong, 2009; Marcoulides, 2018). Bayesian Synthesis is a recently proposed Bayesian approach to data fusion whereby results from the analysis of one dataset are used as prior information in the subsequent analysis of the next dataset (Du et al., 2020; Marcoulides, 2017b), and those results are in turn used as prior information for the analysis of yet another dataset. Datasets of interest are sequentially analyzed in this manner until a final posterior distribution is created, which incorporates information from all datasets of interest.

Conducting data fusion using the Bayesian Synthesis approach thus relies on the sequential updating of estimates as new information from each additional dataset becomes available (for complete technical details on the approach and various recent empirical applications see for example, Fujimoto, Gordon, Peng, & Hofer, 2018; Johnson & Guttmannova, 2019; Marcoulides, 2017a, 2017b, 2018; Preston et al., 2018; Saris & Satorra, 2018). This synthesis notion is expressed using Bayes theorem as

$$P(Unknowns|Data) = \frac{P(Data|Unknowns)P(Unknowns)}{P(Data)} \propto P(Data|Unknowns)P(Unknowns),$$

where $P(Unknowns)$ is the prior probability distribution for the unknown parameters, $P(Data|Unknowns)$ is the conditional probability of the data given the unknown parameters, and $P(Unknowns|Data)$ is the posterior probability distribution for the unknown parameters given our data. Thus when two datasets are fused, the prior information about the unknown parameters can be considered equivalent to a data set that, when merged with the current data, supports the following Bayesian inference $P\left(Unknowns|\, Data_1, Data_2\right) \propto P\left(Data_2|\, Unknowns\right) P\left(Unknowns|\, Data_1\right)$. Here, $P(Unknowns|Data_1)$ is the posterior distribution that resulted from the first analysis where information from $Data_1$ was incorporated with *P(Unknowns)* and then serves as the prior distribution for the present analysis that incorporates the data in $Data_2$. When k datasets are to be fused, the process can be denoted in a general form as $P\left(Unknowns|\, Data_1, \ldots, Data_{k+1}\right) \propto P(Data_{k+1}|\, Unknowns)P\left(Unknowns|\, Data_1, \ldots, Data_k\right)$ with the priors and posterior distribution similarly updated.

A major benefit of the Bayesian Synthesis approach over traditional frequentist approaches to data fusion is the ability to incorporate datasets for which raw data is not available. In this manner, the Bayesian Synthesis approach provides an alternative to the necessity to analyze the raw data and instead uses the estimates and summary statistics from the examined studies to incorporate into the prior information. The approach therefore utilizes point summary estimates of the posterior distributions instead of the actual full posterior distributions as required by a fully Bayesian execution of this Bayesian Synthesis

approach. Bayesian Synthesis thereby enables summary information from published (or unpublished) research to be incorporated as prior information (i.e., an informative prior) for the analysis of another dataset. The sequential use of informative priors that are based on the information in past data provides an extra source of information to estimate model parameters and this additional information can effectively aid in the accuracy of parameters estimation and in the interpretation of results. However, one concern with the Bayesian Synthesis approach is that it heavily relies on updating the information as new data summary statistics become available, therefore the order in which the data are sequentially analyzed may have an impact on the results (Marcoulides, 2017b). Theoretically, in the Bayesian Synthesis approach this should not be a concern due to the conventional Bayesian exchangeability assumption (de Finetti, 1972, 1974), however, Bayesian Synthesis utilizes point summary estimates of the posterior distributions instead of the full posterior distribution as required for a fully Bayesian execution of this approach. While this has the potential to introduce some bias, using point summary estimates of the posterior distributions greatly increases the ease of execution and enables researchers to straightforwardly implement the Bayesian Synthesis approach in standard programs like M*plus* (Muthen & Muthen, 2017). The cost of potentially introducing bias may outweigh the difficulties of incorporating the full distributions in the sequential analysis (Marcoulides, 2017b).

To address the concern about the order of the datasets being analyzed, Marcoulides (2017b) examined the exchangeability assumption and found that the order of analysis did not meaningfully impact the final data fusion results. Similar conclusions regarding exchangeability were also recently suggested by Miocevic, Levy, and Savord (2020). One limitation with these conclusions is that they were based on analyzed datasets that were from similarly-sized large samples. Therefore, it is still unknown whether the order of datasets matters when the datasets being fused each have substantially different sample sizes (as is quite common with empirical data). For example, it may be that beginning the Bayesian Synthesis approach with the analysis of a large dataset produces a substantially biased final posterior distribution when the other sequentially analyzed datasets are much smaller, or vice versa.

In this study, we focus on this unexamined scenario in which there are multiple datasets of both small and large sample sizes. Our main question is whether the order in which datasets are incorporated in the Bayesian Synthesis process will impact the results when one dataset is substantially larger than the rest. To evaluate the performance of Bayesian Synthesis with varied sample sizes, results from simulated data with known population values will be examined under a variety of design study conditions. We conclude with a discussion of the results, implications of the findings, and suggestions for further research.

2 Methods

2.1 Monte Carlo Data Simulation

In order to systematically evaluate the exchangeability of datasets with varying sample sizes in the Bayesian Synthesis approach, simulated data using Monte Carlo techniques were analyzed under a variety of longitudinal data design conditions. All simulated data were generated using R (R Development Core Team, 2010), and analyzed in M*plus* (Muthen & Muthen, 2017) through R using the MplusAutomation package (Hallquist & Wiley, 2014). The analyses were specified to use the Gibbs (PX1) algorithm with a minimum of 50,000 iterations, using the Potential Scale Reduction (PSR) convergence criteria of 1.05 (Gelman et al., 2014), a median summarized posterior, 250 replications, and two chains without thinning.

The simulated data were modeled after the Marcoulides and Grimm (2017) study, which analyzed six longitudinal studies measuring students' mathematics ability. These studied six datasets varied in their sample size, timing, and number of measurement occasions. As Marcoulides and Grimm (2017) showed that mathematics ability increased as children got older, we used the linear growth model to generate the simulated data. In this model, individuals (n) are measured on their math abilities (y) across multiple measurement occasions (t), using the data generation model

$$y_{tn} = \eta_{0n} + \left(\frac{t - k_1}{k_2}\right)\eta_{1n} + e_{tn}, \tag{1}$$

where η_{0n} is an individual's latent intercept when time t equals zero, η_{1n} is an individual's latent slope when time t equals zero, k_1 and k_2 are functional variables to center the intercept and scale the slope, and e_{tn} is the residual at time t for individual n. In this model, we assume the two latent variables follow a multivariate normal distribution, $[\eta_{0n},\ \eta_{1n}]' \sim N(\boldsymbol{\beta},\ \boldsymbol{\Psi})$; and the residuals are assumed to follow a normal distribution, $e_{tn} \sim N(0,\ \sigma_e^2)$, with mean 0 and constant variance.

In all simulation conditions, the latent intercept and slope means were fixed at $\beta_{Intercept}$= -2 and $\beta_{Slope} = 0.4$, and the residual variance σ_e^2 was fixed at 0.10 to reflect small amounts of residuals. Asparouhov and Muthén (2010) and McNeish (2016) indicated that, compared to other parameters of the linear growth model, the choice of prior distributions for the variance-covariance matrix ($\boldsymbol{\Psi}$) in this model is extremely important. Therefore, we additionally varied the $\boldsymbol{\Psi}$ matrix to reflect small, medium, and large magnitudes of their variances, and zero and small magnitudes of their covariances, resulting in the following three variance–covariance matrices: $\boldsymbol{\Psi_1} = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$, $\boldsymbol{\Psi_2} = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$, and $\boldsymbol{\Psi_3} = \begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{bmatrix}$. A summary of these population parameters is presented in Table 1 below. Figure 1 provides an illustrative display of data score plots for

one simulated sample of N = 50 observations based on these three variance-covariance matrices for a simulated data design condition comprised of observations obtained across 3 assessment occasions taken every 5 years starting at age 5.

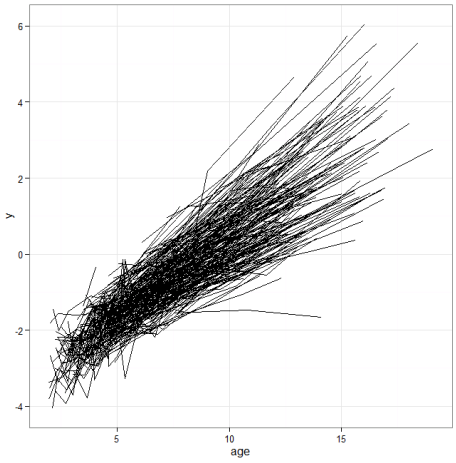
Table 1: Population matrices and covariance parameters.

$\boldsymbol{\Psi}$	$\boldsymbol{\beta_{Intercept}}$	$\boldsymbol{\beta_{Slope}}$	σ_e^2
$\begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$	-2.00	0.40	0.10
$\begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$	-2.00	0.40	0.10
$\begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{bmatrix}$	-2.00	0.40	0.10

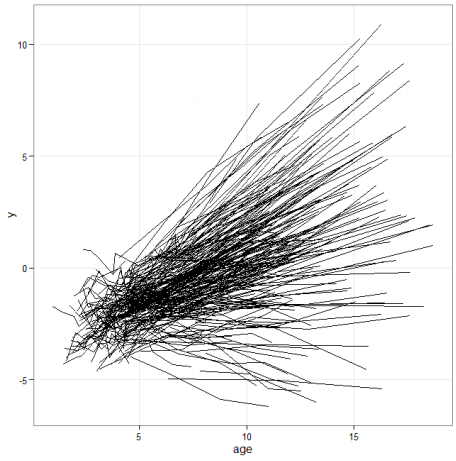
For each variance-covariance matrix condition considered, six datasets were simulated. Because the impact of sample size on exchangeability may also depend on how well the particular sample matches the population of interest, the six datasets were varied with respect to the number of assessments, years between assessments, the age of participants' first assessment, and the sample size. The data patterns for these six datasets are presented in Table 2 and are meant to reflect the full growth trajectory (i.e., across the full age range of interest) with early and late age ranges, as well as large and small numbers of observations across different numbers of assessment occasions. As indicated, the sample sizes for these six datasets were also varied across each $\boldsymbol{\Psi}$ matrix condition, such that each of these datasets were simulated to have a sample size of 1000 and incorporated into the Bayesian synthesis approach as the first dataset, randomly varying the order of the remaining 5 datasets each with sample size of 50, and then again as the last dataset, randomly varying the order of the preceding 5 datasets. These different sample sizes of 50 and 1,000 were selected to reflect small and large sample studies that are commonly encountered in longitudinal data analyses (McNeish, 2016; Paxton, Curran, Bollen, Kirby, & Chen, 2001).

The different specifications for the simulated data presented in detail in Table 2 resulted in a total of 36 different simulated data conditions (3 $\boldsymbol{\Psi}$ matrices, 6 data patterns, and 2 fusing sequences). For example, for the first $\boldsymbol{\Psi_1}$ matrix condition, if dataset 1 (measured 3 times with five years between assessments, starting at age 5) was simulated to have 1000 individuals, the other 5 datasets were then simulated to have a sample size of 50 observations. Thus, the Bayesian Synthesis approach begins with the analysis of dataset 1 with 1,000 observations and produces a posterior point estimate that is then used as the informative prior for the analysis of say dataset 2 with 50 individuals. This process contin-

$$\Psi_1 = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$$



$$\Psi_2 = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$$



$$\Psi_3 = \begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{bmatrix}$$

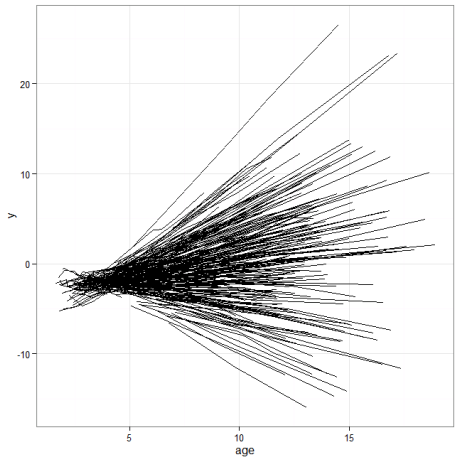


Figure 1: Illustrative data score plots for N = 50 for the first data design.

Table 2: List of the patterns of measurement occasions to be used in the simulated data

Dataset	Number of Assessments	Years between Assessments	StaringAge
1	3	5	5
2	10	1	5
3	10	0.5	2.5
4	10	0.5	10
5	3	0.5	4
6	3	1	11

ues accordingly for each of the remaining datasets until a final posterior point estimate is produced that incorporates information from all 6 datasets. As noted above, this same process is also conducted again but instead with the 1000 observations utilized as the last dataset while randomly varying the order of the preceding 5 datasets each with 50 observations (see additional details below).

2.2 The Bayesian Synthesis Approach

After simulating the datasets according to the conditions described above, we conducted data fusion using the Bayesian Synthesis approach. We began this sequential data integration process with non-informative priors for the analysis of the first dataset because, according to Asparouhov and Muthén (2010), these initial non-informative priors should not introduce bias, even in small sample sizes. The rationale for using non-informative priors also follows recommendations provided by Gelman et al. (2014) to "... let the data speak for themselves, so that inferences are unaffected by information external to the current data (pg. 51)". This is because "the information about model parameters contained in the data will far outweigh any reasonable prior probability specification" (Gelman et al., 2014). Similar recommendations concerning the use of non-informative priors for estimation in growth curve analyses were also provided by Liu, Zhang, and Grimm (2016). So, unless dependable prior information about the range of possible values that model parameters might take is available, beginning the sequential data integration process with non-informative priors appears to be preferable to simply taking a guess at the values for the priors and using informative inaccurate priors, which are known to yield less accurate estimates than non-informative priors (e.g., Shi & Tong, 2017, 2018).

Thus, the Bayesian Synthesis approach begins by using non-informative priors as parameters for the first data set. This implies that for the intercept and slope means a Normal prior of the form N(mean, variance) is used with N(0, 10^{10}), which is the default for a non-informative Normal prior in M*plus* (Muthen & Muthen, 2002). Then, for the parameters in the $\mathbf{\Psi}$ matrix the Inverse Wishart prior IW(0,-3) is used, which is also the default non-informative prior in *Mplus* (Muthen & Muthen, 2002). This prior is of the general form IW(S, d), where

d is the pseudo-sample size and S is the scale matrix $\begin{bmatrix} d(\sigma^2_{Intercept}) & d(\sigma_{IS}) \\ d(\sigma_{IS}) & d(\sigma^2_{Slope}) \end{bmatrix}$, with the estimated intercept ($\sigma^2_{Intercept}$) and slope (σ^2_{Slope}) variances and their covariance (σ_{IS}). Finally, the Inverse Gamma IG(-1, 0) for the residual variance, which is also the default non-informative prior in *Mplus* (Muthen & Muthen, 2017). The Inverse Gamma is of the general form IG(α, β), in which $\alpha = \nu_0/2$ and $\beta = \nu_0\sigma_0^2/2$, and where σ_0^2 can be interpreted as the best estimate of the variance and ν_0 can be interpreted as a pseudo-sample size. Upon analyzing the first data set based on the non-informative priors, posterior point summary estimates for the $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, and σ_e^2 parameters are then sequentially substituted into the respective priors for the next data analysis and the pseudo-sample size of the current data set is then changed by the sample size of the previous data set. This process then makes the priors in the subsequent analyses informative priors and continues sequentially until the sixth and final data set is analyzed and the final posterior distribution and point estimates produced. The simulation results for these analyses are presented in Tables 3a through 8a.

To further investigate the extent to which the use of non-informative versus informative priors at the outset of Bayesian Synthesis might also induce some sort of bias into the obtained results, we also performed all sequential data integration processes using informative priors for the analysis of the first dataset. Given that in practice it may be unlikely for a researcher to have exactly accurate prior information regarding the parameter values, we followed the recommendations of Depaoli (2014) and Finch and Miller (2019) to use "informative" priors as those in which the specified priors correspond to the estimated maximum likelihood growth parameters for each model. Using these estimated values as informative priors, the analyses were then repeated using the 36 different specified simulation data conditions (3 $\boldsymbol{\Psi}$ matrices, 6 data patterns, and 2 fusing sequences) and are presented in Table 3b to Table 8b.

2.3 Parameter Evaluation

The final posterior point summary estimates obtained for each of the 36 different simulated data conditions were evaluated in terms of their raw bias ($B\,(\hat{y})$), relative bias ($RB\,(\hat{y})$), accuracy ($RMSE\,(\hat{y})$), and efficiency ($\textit{Efficiency}(\hat{y})$. These criteria were selected based on past research on dependable parameter evaluation benchmarks (Bandalos & Gagne, 2012; Bandalos & Leite, 2013). These criteria were computed based on the following formulas:

$$B\,(\hat{y}) = \frac{\sum_{r=1}^{R}(\hat{y}_r - y)}{R} \tag{2}$$

$$RB\,(\hat{y}) = \left(\frac{\hat{y} - y}{y}\right) \times 100 \tag{3}$$

$$RMSE\,(\hat{y}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{y}_r - y)^2}{R-1}} \tag{4}$$

$$\textit{Efficiency}\,(\hat{y}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{y}_r - \overline{\hat{y}})^2}{R-1}} \tag{5}$$

where R represents the total number of simulation replications, which is 250 in our study; $\hat{y}$ is the estimated parameter; y is the known population value for our simulation; and $\overline{\hat{y}}$ is the average parameter estimate.

Raw bias, accuracy, and efficiency in essence evaluate the average deviation, the square root of the average deviation, and the variability of the final posterior distribution means, respectively. Nonzero positive or negative values of raw bias indicates overestimation or underestimation respectively. Lower values of accuracy correspond to more precise estimates of the parameters, or estimates of parameters that exhibit a smaller range of error (Bandalos & Gagne, 2012). Values closer to zero correspond to more efficient estimates of the parameters. In other words, smaller values correspond to a smaller range of variability, or higher consistency of estimation. In contrast to these criteria, the magnitude of relative bias is expressed on the percentage scale and indicates the percent deviation of the estimate from the population parameter. This measure is ideal for comparisons of the magnitude of bias across different design conditions (Muthen & Muthen, 2002). To evaluate relative bias, it has been suggested that values less than 5% reflect ignorable bias, values between 5% to 10% indicate moderate bias, and values larger than 10% are considered substantial bias (Muthen & Muthen, 2002). Because multiple simulation replications were conducted (in this case 250 replications were analyzed for all linear growth models examined), average values of relative bias for each evaluated parameter are reported across all the replications. In general, the values of raw bias, accuracy, and efficiency are typically much harder to unravel, whereas values of relative bias are much easier to interpret; we therefore pay extra attention to disentangling obtained relative bias values in reviewing the crucial findings in our study.

3 Results

The simulation results for examining the performance of the exchangeability principle applied within the Bayesian Synthesis approach are presented in Tables 3a through 8b. Within each table, the results are organized based on the magnitude of the variance-covariance $\mathbf{\Psi}$ matrix (in order of small, medium, and large magnitudes) and the order of data fusion (i.e., with the large dataset fused first versus last). The computed parameter estimates reported include the latent variable means $\beta_{Intercept}$ and β_{Slope}, their variances and covariances in terms of $\sigma^2_{Intercept}$, σ^2_{Slope}, σ_{IS}, and the residual variance σ^2_e. Each table presents findings based on the designated evaluation criteria (raw bias, relative bias, accuracy, and efficiency) for each of the parameter estimates across the examined data

design conditions with both non-informative and informative priors. Except for estimates of $\beta_{Intercept}$, positive values of the criteria indicate positive bias or overestimation and negative values indicate negative bias or underestimation. Because $\beta_{Intercept}$ was fixed in the simulation conditions at a negative value, obtaining a negative bias actually corresponds to overestimation and a positive bias corresponds to underestimation.

The results for the first data design condition with specified covariance matrices $\Psi_1 = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$, $\Psi_2 = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$, and $\Psi_3 = \begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{bmatrix}$ are presented in Table 3a. The first data design condition comprised observations obtained across 3 assessment occasions taken every 5 years starting at age 5. This dataset reflected the feature of large breadth with small numbers of assessments while covering different age ranges. The reported results correspond to those obtained when (i) a sample size of 1000 observations is incorporated as the first dataset into the Bayesian synthesis approach while randomly varying the order of the remaining 5 datasets each with a sample size of 50, and (ii) a sample size of 1000 observations is incorporated into the Bayesian synthesis approach as the last dataset while randomly varying the order of the preceding 5 datasets each with a sample of size 50. To make this distinction clear, the results are labelled in this and all subsequent tables as FIRST and LAST for each reported evaluation criterion.

In general, the obtained values for the raw bias, accuracy, and efficiency criteria are larger when the larger dataset is fused last instead of first in the Bayesian synthesis approach. However, these values are all still very close to zero, indicating precise and efficient estimates of the parameters. These obtained bias values also appear to be similar regardless of which specified variance-covariance matrix is examined. Looking for example at the estimated intercept variance ($\sigma^2_{Intercept}$) for the first dataset condition in Table 3a, some patterns of results can be detected. For instance, the obtained values for the raw bias (.0011, .0048, and .0029), accuracy (.0133, .0331, and .0212), and efficiency (.0133, .0327, .0210) when the large dataset was fused first are negligible. However, when the large dataset is fused last, the estimated intercept variance values increase (though are still rather close to zero) for the raw bias (.1244, .1300, and .1286), the accuracy (.2664, .2828, and .2776), and efficiency (.2355, .2511, .2459). Similar patterns of results are observed when informative priors are used to analyze the large data set first (see Table 3b). As indicated previously, because the values of raw bias, accuracy, and efficiency are typically more challenging to unravel, we instead pay extra attention to disentangling the obtained relative bias values as these are generally easier to interpret.

When examining the relative bias criterion under the three variance-covariance matrixes, some interesting patterns of results are again revealed. Specifically, with initial non-informative priors it can be seen that the relative bias for the estimated intercept variance ($\sigma^2_{Intercept}$) increases from ignorable sizes (0.546%, 0.692%, and 0.740%, respectively) when the large dataset was fused first into substantially biased values (62.208%, 18.572%, and 32.161%, respectively) when the large dataset was fused last. The estimates of the slope variance (σ^2_{Slope}) also

showed somewhat similar patterns of results, but different magnitudes when fusing the larger sample dataset last. The variances changed from ignorable biased (0.720%, 0.880%, and 0.802%, respectively) when the large dataset was fused first, to substantially biased values (58.12%) for the small magnitude covariance matrix (Ψ_1) and moderately biased (8.31%, 6.336%) for medium and large magnitude covariance matrixes (Ψ_2 and Ψ_3) when the larger dataset was fused last. Another sizable amount of relative bias was also observed when examining the magnitude of the intercept slope covariance value (σ_{IS}) for the Ψ_2 variance-covariance matrix, shifting from ignorable bias (0.384%) when the large dataset was fused first to substantially biased (-12.984%) when the larger dataset was fused last.

In contrast, when initial informative priors are used, the relative bias for the residual variance when the large sample is incorporated into the Bayesian synthesis approach as the last dataset were found to be moderate across the three covariance matrix conditions (7.248 %, 5.244%, and 5.108%). Additionally, the relative bias for the estimated intercept ($\sigma^2_{Intercept}$) and slope (σ^2_{Slope}) variances increased from ignorable sizes when the large dataset was fused first into substantially biased for the first covariance matrix condition (12.01% and 16.52% respectively) when the large dataset was fused last.

The observed relative bias for the variance of the intercept, the slope, and to a lesser degree the intercept-slope covariance in this first data design condition highlight the importance that the order of fusing the datasets can play in the estimation of parameters when implementing Bayesian Synthesis strategies. Interestingly, this finding is not in line with past research that has suggested that the order of data fusion does not meaningfully impact the final posterior distribution results (Marcoulides, 2017b; Miocevic et al., 2020). When the data sets being fused are of differing sizes (50 vs. 1,000), ending with the fusion and analysis of a large dataset can in fact produce a substantially biased final posterior distribution when the other sequentially analyzed datasets are much smaller. However, these results are only discernable when using the measure of relative bias, they do not appear sizeable when examining the values of raw bias, accuracy, or efficiency.

The results for the second simulated data design condition for the variance-covariance matrices Ψ_1, Ψ_2, and Ψ_3 are presented in Table 4a. The second data design condition comprised of observations across 10 assessment occasions, measured every year, starting from age 5. This simulated condition reflected the feature of large breadth of measurement years covering different age ranges. As described previously, the reported results correspond to those obtained when a sample size of 1000 observations is incorporated as the first dataset and again as the last dataset while randomly varying the order of the other 5 datasets each with 50 observations. The results are similarly labelled as FIRST and LAST for each evaluation criterion examined.

In general, the obtained values for the raw bias, accuracy, and efficiency criteria reflect similar results to those observed for the first simulated data design condition. Looking for example at the estimated intercept variance ($\sigma^2_{Intercept}$)

Table 3a: Data Condition 1 Using Initial Non-Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
σ_e^2	0.0005[a]	-0.0009	0.4840[a]	-0.9440	0.0033[a]	0.0055	0.0033[a]	0.0054
	0.0006[b]	-0.0002	0.5600[b]	-0.2200	0.0033[b]	0.0042	0.0033[b]	0.0042
	0.0005[c]	-0.0003	0.5000[c]	-0.2880	0.0034[c]	0.0053	0.0033[c]	0.0053
σ_{IS}	0.0001	-0.0154	−−	−−	0.0019	0.0323	0.0019	0.0284
	0.0002	0.0065	0.3840	**-12.984**	0.0083	0.0389	0.0083	0.0384
	0.0000	0.0040	0.3520	0.1960	0.0140	0.0614	0.0140	0.0614
$\beta_{Intercept}$	0.0007	0.0016	-0.0334	-0.0776	0.0163	0.0172	0.0163	0.0171
	0.0011	0.0020	-0.0570	-0.0986	0.0267	0.0267	0.0267	0.0266
	0.0007	0.0015	-0.0354	-0.0740	0.0212	0.0209	0.0212	0.0208
β_{Slope}	-0.0001	-0.0001	-0.0220	-0.0130	0.0034	0.0034	0.0034	0.0034
	0.0001	0.0002	0.0130	0.0450	0.0091	0.0092	0.0091	0.0092
	0.0006	0.0008	0.1430	0.2000	0.0181	0.1790	0.0181	0.0179
$\sigma^2_{Intercept}$	0.0011	0.1244	0.5460	**62.208**	0.0133	0.2664	0.0133	0.2355
	0.0048	0.1300	0.6920	**18.573**	0.0331	0.2828	0.0327	0.2511
	0.0029	0.1286	0.7370	**32.161**	0.0212	0.2776	0.0210	0.2459
σ^2_{Slope}	0.0001	0.0058	0.7200	**58.120**	0.0006	0.0107	0.0006	0.0090
	0.0009	0.0083	0.8800	**8.312**	0.0043	0.0189	0.0042	0.0169
	0.0032	0.0253	0.8020	**6.336**	0.0168	0.0589	0.0165	0.0531

Note: [a] Denotes results for covariance matrix Ψ_1, [b] denotes results for covariance matrix Ψ_2, and [c] denotes results for covariance matrix Ψ_3. The relative bias for estimates of the intercept-slope covariance for Ψ_1 cannot be computed, as the population value was zero. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters values negative bias corresponds to underestimation and positive bias corresponds to overestimation. Bolded values indicate moderate or substantial bias.

Table 3b: Data Condition 1 Using Initial Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
	0.0003[a]	0.0072	0.3280[a]	**7.2480**	0.0031[a]	0.0169	0.0031[a]	0.0152
σ_e^2	0.0004[b]	0.0052	0.3640[b]	**5.2440**	0.0032[b]	0.0220	0.0031[b]	0.0214
	0.0003[c]	0.0051	0.3240[c]	**5.1080**	0.0031[c]	0.0172	0.0031[c]	0.0164
	-0.0001	-0.0050	−−	−−	0.0020	0.0233	0.0020	0.0228
σ_{IS}	-0.0003	-0.0025	-0.6880	-4.9920	0.0085	0.0380	0.0085	0.0379
	-0.0006	-0.0098	-0.2820	-4.9060	0.0148	0.0615	0.0148	0.0607
	0.0000	0.0004	-0.0016	-0.0206	0.0156	0.0166	0.0156	0.0166
$\beta_{Intercept}$	-0.0003	0.0009	0.0130	-0.0468	0.0254	0.0254	0.0254	0.0254
	-0.0004	0.0011	0.0200	-0.0528	0.0202	0.0217	0.0202	0.0217
	-0.0002	0.0000	-0.0420	0.0010	0.0033	0.0036	0.0033	0.0036
β_{Slope}	-0.0002	0.0003	-0.0490	0.0700	0.0093	0.0094	0.0093	0.0094
	-0.0004	0.0008	-0.0940	0.1930	0.0185	0.0196	0.0185	0.0196
	0.0007	0.0240	0.3260	**12.0100**	0.0136	0.1740	0.0136	0.1723
$\sigma^2_{Intercept}$	0.0014	-0.0103	0.1994	-1.4657	0.0337	0.2292	0.0337	0.2290
	0.0015	0.0128	0.3690	3.1940	0.0216	0.1890	0.0216	0.1886
	0.0000	0.0017	-0.3200	**16.520**	0.0006	0.0066	0.0006	0.0064
σ^2_{Slope}	0.0002	0.0006	0.1920	0.6080	0.0042	0.0153	0.0042	0.0152
	0.0004	0.0009	0.0960	0.2210	0.0159	0.0416	0.0159	0.0416

Note: Same as Table 3a.

for the first dataset condition in Table 4a, the same patterns of results can be detected. For instance, the obtained values for the raw bias (.0064, .0219, and .0127), accuracy (.0196, .0685, and .0396), and efficiency (.0186, .0649, .0375) when the large dataset was fused first are negligible. However, when the large dataset is fused last, the estimated intercept variance values increase (though are still rather close to zero) for the raw bias (.0688, .0756, and .0721), accuracy (.1444, .1521, and .1474), and efficiency (.1268, .1319, .1285).

Focusing on the measure of relative bias with initial non-informative priors, the biased values are observed when comparing estimates to the true population values when the larger dataset was fused last in the Bayesian synthesis approach and regardless of the specified variance-covariance matrix examined. Specifically, one can see that the relative bias for the estimated intercept variance ($\sigma^2_{Intercept}$) increases from ignorable sizes (3.192%, 3.129%, and 3.185%, respectively) when the large dataset was fused first into substantially biased values (34.414%, 10.804%, and 18.013%, respectively) when the large dataset is fused last. The estimates of the slope variance (σ^2_{Slope}) also showed somewhat similar patterns of results, but with somewhat different magnitudes when fusing the larger sample dataset last. The variances changed from ignorable bias (1.52%, 1.56%, and 1.89%, respectively) when the large dataset was fused first, to substantially biased values (33.92%) for the small magnitude covariance matrix (Ψ_1) to moderately biased (5.86%) for medium magnitude covariance matrix (Ψ_2) to ignorable (4.53%) for the large magnitude covariance matrix (Ψ_3) when the larger dataset was fused last. A sizable amount of relative bias was again observed when examining the magnitude of the intercept slope covariance value (σ_{IS}) for the Ψ_2 variance-covariance matrix, shifting from ignorable bias (1.84%) when the large dataset was fused first to substantially biased (-10.696%) when the larger dataset was fused last. Table 4b displays the results for when initial informative priors are used in the data fusion process. Here, moderate bias is only observed for the estimates of the intercept variance ($\sigma^2_{Intercept}$) and slope variance (σ^2_{Slope}) for the first covariance matrix condition (Ψ_1) when the large dataset is analyzed last (7.706% and 6.96% respectively).

The findings obtained under the second simulated data design condition once again highlight the importance that the order of fusing the datasets plays in the estimation of parameters when implementing Bayesian Synthesis strategies. Fusing data in which a larger dataset is fused last can produce a substantially biased final posterior distribution when the other sequentially analyzed datasets are much smaller.

It is important to note that these same patterns of results were also observed for the both the third (see Table 5a) and the fourth (see Table 6a) simulated data design conditions using both non-informative and informative priors for the initial dataset in the data fusion process (see Tables 5b and 6b). The third data design condition encompassed a longitudinal study with 10 assessment occasions, measured every 6 months, starting from age 2.5. This data design was meant to reflect small breadth studies that start at an early age range but with large numbers of observations. The fourth data design condition also covered

Table 4a: Data Condition 1 Using Initial Non-Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
σ_e^2	0.0002[a]	0.0000	-0.208[a]	0.0080	0.0051[a]	0.0016	0.0051[a]	0.0016
	0.0002[b]	0.0001	-0.232[b]	0.1360	0.0051[b]	0.0016	0.0051[b]	0.0016
	0.0002[c]	0.0001	-0.236[c]	0.1160	0.0051[c]	0.0016	0.0051[c]	0.0016
σ_{IS}	0.0001	-0.0093	–	–	0.0029	0.0196	0.0029	0.0173
	0.0009	-0.0053	1.8400	**-10.696**	0.0172	0.0274	0.0171	0.0269
	0.0045	-0.0012	2.274	-0.592	0.0297	0.0416	0.0293	0.0416
$\beta_{Intercept}$	-0.0019	-0.0010	0.0940	0.0506	0.0151	0.0152	0.0150	0.0152
	-0.0034	-0.0024	0.1694	0.1224	0.0260	0.0256	0.0257	0.0255
	-0.0024	-0.0019	0.1194	0.0930	0.0206	0.0202	0.0204	0.0201
β_{Slope}	0.0001	0.0001	0.0250	0.0350	0.0031	0.0032	0.0031	0.0032
	0.0000	0.0003	0.0006	0.0750	0.0089	0.0089	0.0089	0.0089
	0.0004	0.0002	-0.1010	-0.0540	0.0177	0.0171	0.0177	0.0171
$\sigma^2_{Intercept}$	0.0064	0.0688	3.1920	**34.414**	0.0196	0.1444	0.0186	0.1268
	0.0219	0.0756	3.1286	**10.804**	0.0685	0.1521	0.0649	0.1319
	0.0127	0.0721	3.1850	**18.013**	0.0396	0.1474	0.0375	0.1285
σ^2_{Slope}	0.0002	0.0034	1.5200	**33.920**	0.0010	0.0060	0.0010	0.0050
	0.0016	0.0059	1.5600	**5.860**	0.0090	0.0123	0.0089	0.0108
	0.0076	0.0174	1.8920	4.531	0.0363	0.0383	0.0355	0.0342

Note: Same as Table 3a.

Table 4b: Data Condition 2 Using Initial Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
	0.0000[a]	0.0010	-0.0240[a]	1.0080	0.0048[a]	0.0036	0.0048[a]	0.0035
σ_e^2	0.0000[b]	0.0004	0.0440[b]	0.3760	0.0049[b]	0.0025	0.0049[b]	0.0025
	0.0001[c]	0.0003	0.0600[c]	0.3200	0.0048[c]	0.0020	0.0048[c]	0.0020
	0.0000	-0.0016	−−	−−	0.0031	0.0138	0.0031	0.0137
σ_{IS}	0.0009	0.0021	1.7600	4.1680	0.0183	0.0278	0.0182	0.0277
	0.0040	0.0030	2.0120	1.5140	0.0307	0.0378	0.0304	0.0376
	-0.0001	0.0000	0.0036	-0.0006	0.0147	0.0157	0.0147	0.0157
$\beta_{Intercept}$	-0.0005	-0.0003	0.0252	0.0144	0.0255	0.0263	0.0255	0.0263
	-0.0004	0.0001	0.0224	-0.0034	0.0198	0.0219	0.0198	0.0219
	-0.0002	0.0000	-0.0580	-0.0120	0.0036	0.0035	0.0036	0.0035
β_{Slope}	-0.0007	-0.0003	-0.1690	-0.0710	0.0100	0.0097	0.0100	0.0097
	-0.0010	-0.0007	-0.2580	-0.1760	0.0194	0.0196	0.0194	0.0196
	0.0063	0.0154	3.1580	**7.7060**	0.0192	0.0889	0.0181	0.0876
$\sigma^2_{Intercept}$	0.0225	0.0129	3.2171	1.8446	0.0672	0.1236	0.0633	0.1229
	0.0133	0.0173	3.3130	4.3160	0.0385	0.1001	0.0361	0.0985
	0.0001	0.0007	1.1200	**6.9600**	0.0010	0.0043	0.0010	0.0043
σ^2_{Slope}	0.0015	0.0001	1.4920	0.0640	0.0089	0.0114	0.0088	0.0114
	0.0067	0.0032	1.6710	0.7950	0.0376	0.0340	0.0370	0.0339

Note: Same as Table 3a.

10 assessment occasions, measured every six months, but starting instead from age 10. This data design represented the feature of small breadth with a large number of observations that covered a late age range. Given the similarity of these observed bias findings, we focus next on examining the results for just the fifth and sixth simulated data design conditions.

Table 5a: Data Condition 3 Using Initial Non-Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
σ_e^2	0.0003[a]	-0.0001	0.324[a]	-0.0064	0.0054[a]	0.0016	0.0054[a]	0.0016
	0.0003[b]	0.0002	0.284[b]	0.1520	0.0055[b]	0.0016	0.0055[b]	0.0016
	0.0002[c]	0.0001	0.244[c]	0.1280	0.0055[c]	0.0016	0.0055[c]	0.0016
σ_{IS}	-0.0003	-0.0078	–	–	0.0031	0.0159	0.0031	0.0138
	0.0002	-0.0050	0.408	**-10.008**	0.0177	0.0229	0.0177	0.0223
	0.0037	-0.0018	1.846	-0.8820	0.0300	0.0331	0.0297	0.0331
$\beta_{Intercept}$	-0.0017	-0.0008	0.0862	0.0406	0.0143	0.0143	0.0142	0.0143
	-0.0031	-0.0025	0.1526	0.1226	0.0253	0.0253	0.0252	0.0252
	-0.0022	-0.0020	0.1080	0.0992	0.0198	0.0198	0.0197	0.0197
β_{Slope}	0.0001	0.0001	0.0200	0.0270	0.0034	0.0036	0.0034	0.0036
	-0.0001	0.0003	-0.0170	0.0780	0.0091	0.0090	0.0091	0.0090
	-0.0006	-0.0001	-0.1530	-0.0270	0.0178	0.0173	0.0178	0.0173
$\sigma^2_{Intercept}$	0.0071	0.0530	3.5620	**26.518**	0.0195	0.1078	0.0182	0.0938
	0.0212	0.0600	3.0354	**8.5703**	0.0673	0.1161	0.0639	0.0993
	0.0130	0.0563	3.2560	**14.078**	0.0390	0.1117	0.0367	0.0964
σ^2_{Slope}	0.0003	0.0029	2.8400	**28.680**	0.0011	0.0050	0.0001	0.0041
	0.0029	0.0050	2.8960	**5.048**	0.0095	0.0102	0.0090	0.0088
	0.0108	0.0145	1.2700	3.628	0.0371	0.0317	0.0355	0.0282

Note: Same as Table 3a.

The results for the fifth simulated data design condition for the variance-covariance matrices Ψ_1, Ψ_2, and Ψ_3 are presented in Table 7a. This simulated data design condition comprised of observations taken across 3 assessment occasions, measured every six months, starting from age 4. This simulated condition reflected the feature of early age range of development in a small breadth of measurement years. The results presented in Table 7a again correspond to those obtained when a sample size of 1000 observations is incorporated as the first dataset and then as the last dataset while randomly varying the order of the other 5 datasets each with 50 observations.

Table 5b: Data Condition 3 Using Initial Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
	0.0004[a]	0.0009	0.4360[a]	0.9120	0.0050[a]	0.0032	0.0050[a]	0.0031
σ_e^2	0.0004[b]	0.0003	0.4080[b]	0.2720	0.0051[b]	0.0021	0.0050[b]	0.0021
	0.0005[c]	0.0002	0.4680[c]	0.2120	0.0049[c]	0.0018	0.0049[c]	0.0018
	0.0000	-0.0006	−−	−−	0.0032	0.0119	0.0032	0.0119
σ_{IS}	0.0016	0.0043	3.2080	**8.6160**	0.0186	0.0244	0.0185	0.0240
	0.0063	0.0066	3.1480	3.3060	0.0310	0.0325	0.0304	0.0318
	0.0002	0.0002	-0.0088	-0.0106	0.0140	0.0146	0.0140	0.0146
$\beta_{Intercept}$	-0.0002	-0.0006	0.0078	0.0294	0.0253	0.0255	0.0253	0.0255
	0.0001	-0.0006	-0.0040	0.0276	0.0194	0.0205	0.0194	0.0205
	-0.0002	0.0000	-0.0450	-0.0090	0.0037	0.0039	0.0037	0.0039
β_{Slope}	-0.0003	-0.0004	-0.0830	-0.1000	0.0103	0.0097	0.0103	0.0097
	-0.0006	-0.0013	-0.1570	-0.3240	0.0197	0.0192	0.0197	0.0192
	0.0088	0.0126	4.3800	**6.3160**	0.0208	0.0703	0.0189	0.0692
$\sigma^2_{Intercept}$	0.0298	0.0169	4.2560	2.4177	0.0726	0.1027	0.0662	0.1013
	0.0177	0.0166	4.4180	4.1420	0.0419	0.0815	0.0380	0.0798
	0.0002	0.0006	1.7600	**6.0800**	0.0011	0.0038	0.0011	0.0037
σ^2_{Slope}	0.0030	0.0014	3.0040	1.3720	0.0096	0.0098	0.0091	0.0097
	0.0129	0.0091	3.2320	2.2650	0.0382	0.0306	0.0360	0.0293

Note: Same as Table 3a.

Table 6a: Data Condition 4 Using Initial Non-Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
σ_e^2	0.0004[a]	-0.0002	0.380[a]	-0.2040	0.0052[a]	0.0017	0.0052[a]	0.0017
	0.0003[b]	0.0000	0.304[b]	0.0080	0.0053[b]	0.0016	0.0053[b]	0.0016
	0.0003[c]	-0.0001	0.304[c]	-0.1320	0.0053[c]	0.0017	0.0053[c]	0.0017
σ_{IS}	-0.0001	-0.0079	–	–	0.0035	0.0158	0.0035	0.0137
	0.0004	-0.0053	0.832	**-10.544**	0.0179	0.0224	0.0179	0.0218
	0.0048	-0.0022	2.382	-1.1160	0.0291	0.0336	0.0287	0.0335
$\beta_{Intercept}$	-0.0012	-0.0004	0.0596	0.0194	0.0195	0.0191	0.0194	0.0191
	-0.0026	-0.0020	0.1318	0.1020	0.0282	0.0285	0.0281	0.0284
	-0.0016	-0.0015	0.0802	0.0728	0.0231	0.0241	0.0230	0.0240
β_{Slope}	0.0001	0.0001	0.0130	0.0190	0.0035	0.0034	0.0035	0.0034
	0.0000	-0.0003	-0.0050	0.0640	0.0090	0.0090	0.0090	0.0090
	-0.0008	0.0000	-0.1980	0.0040	0.0176	0.0175	0.0175	0.0175
$\sigma^2_{Intercept}$	0.0077	0.0535	3.8260	**26.726**	0.0225	0.1107	0.0212	0.0969
	0.0235	0.0605	3.3566	**8.6389**	0.0677	0.1160	0.0634	0.0989
	0.0144	0.0577	3.6030	**14.424**	0.0406	0.1138	0.0379	0.0980
σ^2_{Slope}	0.0003	0.0024	2.5200	**23.760**	0.0011	0.0042	0.0011	0.0034
	0.0030	0.0044	2.9600	4.384	0.0094	0.0090	0.0089	0.0078
	0.0110	0.0125	2.7420	3.115	0.0355	0.0279	0.0338	0.0250

Note: Same as Table 3a.

Table 6b: Data Condition 4 Using Initial Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
σ_e^2	0.0003[a]	0.0002	0.3000[a]	0.2440	0.0051[a]	0.0019	0.0051[a]	0.0019
	0.0004[b]	0.0002	0.3600[b]	0.2040	0.0051[b]	0.0021	0.0051[b]	0.0021
	0.0004[c]	0.0001	0.3520[c]	0.0720	0.0053[c]	0.0018	0.0052[c]	0.0018
σ_{IS}	0.0001	-0.0014	−−	−−	0.0036	0.0113	0.0036	0.0112
	0.0019	0.0025	3.7200	**5.0320**	0.0181	0.0231	0.0180	0.0230
	0.0069	0.0038	3.4380	1.9080	0.0302	0.0317	0.0295	0.0315
$\beta_{Intercept}$	-0.0006	0.0001	0.0292	-0.0038	0.0202	0.0213	0.0202	0.0213
	-0.0010	-0.0006	0.0518	0.0320	0.0291	0.0293	0.0291	0.0293
	-0.0006	-0.0006	0.0302	0.0280	0.0242	0.0264	0.0242	0.0264
β_{Slope}	-0.0001	0.0000	-0.0250	-0.0110	0.0038	0.0041	0.0038	0.0041
	-0.0003	-0.0002	-0.0740	-0.0550	0.0102	0.0101	0.0102	0.0101
	-0.0009	-0.0005	-0.2340	-0.1270	0.0195	0.0197	0.0195	0.0197
$\sigma^2_{Intercept}$	0.0084	0.0162	4.2240	**8.1120**	0.0228	0.0695	0.0212	0.0675
	0.0304	0.0178	4.3463	2.5463	0.0706	0.0990	0.0637	0.0974
	0.0178	0.0189	4.4480	4.7240	0.0418	0.0784	0.0378	0.0761
σ^2_{Slope}	0.0003	0.0007	3.4000	**6.6400**	0.0012	0.0032	0.0011	0.0031
	0.0042	0.0021	4.1560	2.0960	0.0104	0.0092	0.0095	0.0089
	0.0165	0.0099	4.1200	2.4740	0.0399	0.0292	0.0363	0.0275

Note: Same as Table 3a.

In general, the obtained values for the raw bias, accuracy, and efficiency criteria reflect similar results to those observed in the other simulated data design conditions. Biased values were observed when comparing the obtained estimate to the true population values when applying the Bayesian Synthesis approach and regardless of the specified variance-covariance matrix examined. Focusing again on the estimated intercept variance ($\sigma^2_{Intercept}$) for the first dataset condition in Table 7a, the obtained values for the raw bias (.0069, .0212, and .0125), accuracy (.0190, .0643, and .0368), and efficiency (.0177, .0607, .0346) when the large dataset was fused first are negligible. However, when the large dataset is fused last, the estimated intercept variance values increase by a little for the raw bias (.0475, .0546, and .0532), accuracy (.0929, .1059, and .1027), and efficiency (.0798, .0907, .0878).

When examining the relative bias criterion under the three variance-covariance matrixes, some visible biased patterns of results again emerge. In this data design, it can again be seen that the relative bias for the estimated intercept variance ($\sigma^2_{Intercept}$) increases from ignorable sizes (3.472%, 3.033%, and 3.116%, respectively) when the large dataset was fused first into moderately biased and substantially biased values (23.726%, 7.793%, and 13.330%, respectively) when the large dataset was fused last. Interestingly, the estimates of the slope variance (σ^2_{Slope}) showed rather different patterns of results, with sizable magnitudes of relative bias both when fusing the larger sample dataset first and last. In particular, the slope variances displayed substantial bias (18.480%) for covariance matrix (Ψ_1) and ignorable bias (4.320% and 3.574%) for covariance matrixes (Ψ_2 and Ψ_3) when the large dataset was fused first, compared to substantially biased values (18.28%) for the small magnitude covariance matrix (Ψ_1) to moderately biased (6.35%) for medium magnitude covariance matrix (Ψ_2) to ignorable (4.42%) for the large magnitude covariance matrix (Ψ_3) when the larger dataset was fused last. A sizable amount of relative bias was also observed when examining the magnitude of the intercept slope covariance value (σ_{IS}) for the Ψ_2 variance-covariance matrix, shifting from ignorable bias (-2.104%) when the large dataset was fused first to substantially biased (-15.112%) when the larger dataset was fused last. This similar pattern was also observed in Table 7b when initial informative priors were used. Specifically, moderate bias for the intercept variance ($\sigma^2_{Intercept}$) for the first covariance matrix (Ψ_1) condition was observed both when the large sample dataset was fused first and last. In contrast, moderate bias for the estimate of the slope variance (σ^2_{Slope}) was only observed when the large dataset was fused last (6.3855%).

Table 8a presents the results for the final sixth simulated data design condition for the variance-covariance matrices Ψ_1, Ψ_2, and Ψ_3. This simulated data design condition comprised of observations collected across 3 assessment occasions, measured every year, starting from age 11. This simulated condition reflected the feature of a late age range of development in a small breadth of measurement years. The obtained values for the raw bias, accuracy, and efficiency criteria reflect similar results to those observed in other simulated data design conditions. Focusing again on the estimated intercept variance ($\sigma^2_{Intercept}$)

Table 7a: Data Condition 5 Using Initial Non-Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
σ_e^2	0.0001[a]	-0.0005	-0.112[a]	-0.5160	0.0053[a]	0.0032	0.0053[a]	0.0032
	0.0001[b]	-0.0005	-0.052[b]	-0.4640	0.0052[b]	0.0035	0.0052[b]	0.0035
	0.0000[c]	-0.0001	-0.036[c]	-1.0200	0.0052[c]	0.0040	0.0052[c]	0.0039
σ_{IS}	-0.0023	-0.0073	−−	−−	0.0049	0.0142	0.0035	0.0121
	-0.0011	-0.0076	-2.104	**-15.112**	0.0156	0.0197	0.0156	0.0182
	0.0008	-0.0079	0.382	-3.9400	0.0251	0.0262	0.0251	0.0250
$\beta_{Intercept}$	-0.0008	-0.0006	0.0378	0.0312	0.0164	0.0152	0.0163	0.0152
	-0.0011	-0.0010	0.0538	0.4840	0.0263	0.0264	0.0263	0.0264
	-0.0015	-0.0010	0.0764	0.0482	0.0208	0.0214	0.0208	0.0213
β_{Slope}	0.0009	0.0009	0.0232	0.2320	0.0062	0.0065	0.0062	0.0064
	0.0015	-0.0020	0.3810	0.4960	0.0131	0.0132	0.0130	0.0131
	0.0017	0.0028	0.4170	0.7050	0.0207	0.0211	0.0207	0.0209
$\sigma^2_{Intercept}$	0.0069	0.0475	3.4720	**23.726**	0.0190	0.0929	0.0177	0.0798
	0.0212	0.0546	3.0331	**7.7931**	0.0643	0.1059	0.0607	0.0907
	0.0125	0.0532	3.1160	**13.300**	0.0368	0.1027	0.0346	0.0878
σ^2_{Slope}	0.0018	0.0018	**18.480**	**18.280**	0.0029	0.0029	0.0023	0.0023
	0.0043	0.0064	4.3200	**6.353**	0.0100	0.0114	0.0090	0.0095
	0.0143	0.0177	3.5740	4.416	0.0367	0.0344	0.0338	0.0295

Note: Same as Table 3a.

Table 7b: Data Condition 5 Using Initial Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
	0.0008[a]	0.0003	0.8273[a]	0.3052	0.0059[a]	0.0037	0.0059[a]	0.0037
σ_e^2	0.0001[b]	-0.0001	0.1360[b]	-0.0920	0.0046[b]	0.0035	0.0046[b]	0.0035
	0.0001[c]	0.0002	0.0760[c]	0.2000	0.0046[c]	0.0043	0.0046[c]	0.0043
	0.0003	-0.0016	−−	−−	0.0058	0.0110	0.0059	0.0109
σ_{IS}	-0.0007	-0.0012	-1.4160	-2.3200	0.0161	0.0192	0.0161	0.0191
	0.0020	-0.0009	1.0160	-0.4480	0.0257	0.0293	0.0256	0.0293
	0.0016	0.0018	-0.0795	-0.0878	0.0165	0.0155	0.0165	0.0154
$\beta_{Intercept}$	0.0025	0.0026	-0.1234	-0.1288	0.0267	0.0261	0.0266	0.0259
	0.0020	0.0016	-0.1000	-0.0802	0.0211	0.0217	0.0210	0.0216
	-0.0002	-0.0001	-0.0562	-0.0351	0.0068	0.0074	0.0068	0.0074
β_{Slope}	0.0001	0.0001	0.0320	0.0310	0.0140	0.0137	0.0140	0.0137
	0.0007	0.0002	0.1870	0.0600	0.0213	0.0220	0.0213	0.0220
	0.0118	0.0127	**5.8976**	**6.3394**	0.0235	0.0613	0.0204	0.0601
$\sigma^2_{Intercept}$	0.0267	0.0140	3.8080	2.0000	0.0632	0.0824	0.0573	0.0812
	0.0161	0.0144	4.0240	3.5950	0.0368	0.0705	0.0331	0.0690
	0.0000	0.0006	0.0402	**6.3855**	0.0028	0.0022	0.0029	0.0022
σ^2_{Slope}	0.0028	0.0029	2.8080	2.9080	0.0095	0.0099	0.0091	0.0095
	0.0105	0.0091	2.6280	2.2650	0.0361	0.0318	0.0345	0.0305

Note: Same as Table 3a.

in Table 8a, the same patterns of results are evident. For instance, the obtained values for the raw bias (.0038, .0214, and .0251), accuracy (.0552, .0972, and .0613), and efficiency (.0399, .0762, .0559) when the large dataset was fused first are negligible. Similarly, when the large dataset is fused last, the estimated intercept variance values again slightly increase (though are still rather close to zero) for the raw bias (.0145, .0335, and .0206), accuracy (.0357, .0808, and .0529), and efficiency (.0326, .0735, .0487).

When carefully examining the relative bias criterion under the three variance-covariance matrixes, some novel biased results emerge. Specifically, it can be seen in Table 8a that the relative bias for the estimated intercept variance ($\sigma^2_{Intercept}$) ranges from substantial to ignorable to moderately biased (19.020%, 3.054%, and 6.286%, respectively) when the large dataset was fused first, and similarly (7.226%, 4.785%, and 5.162%, respectively) when the large dataset was fused last. The estimates of the slope variance (σ^2_{Slope}) also showed somewhat similar patterns of results, with sizable magnitudes of relative bias both when fusing the larger sample dataset first and last. The slope variances also displayed ranges from substantial bias to ignorable bias (13.20%, 4.63%, and 4.04%, respectively) when the large dataset was fused first, compared to (15.52%, 2.93%, and 1.66%, respectively) when the larger dataset was fused last. A sizable amount of relative bias was also observed when examining the magnitude of the intercept slope covariance value (σ_{IS}) for the Ψ_2 variance-covariance matrix, shifting from ignorable bias (-2.016%) when the large dataset was fused first to substantially biased (-13.76%) when the larger dataset was fused last. Interestingly, this was the only data design condition where moderate bias was observed when the large dataset was analyzed first and informative priors were used to analyze the first dataset. In table 8b, we see that there was moderate bias (7.1478% and 5.5466% respectively) for the estimated intercept ($\sigma^2_{Intercept}$) and slope (σ^2_{Slope}) variance for the first variance-covariance matrix condition (Ψ_1).

These results appear to collectively highlight not only the importance that the order that fusing the datasets can play in the estimation of parameters but also the potential impact that data design characteristics can exert when implementing Bayesian synthesis strategies. It appears that in data design settings with fewer occasions of measurement covering wide range of ages, there can be sizable bias irrespective of whether a larger data set is fused first or last. Additionally, even in instances where there are sufficient assessment settings over a wider age range, the order of the fusing of the data sets can again play a key role. These results would collectively suggest that the order in which datasets are incorporated in the Bayesian Synthesis process do in fact impact the results when one dataset is substantially larger than the rest.

Table 8a: Data Condition 6 Using Initial Non-Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
σ_e^2	0.0009[a]	-0.0003	-0.888[a]	-0.3040	0.0051[a]	0.0032	0.0050[a]	0.0032
	0.0002[b]	-0.0004	-0.024[b]	-0.3920	0.0051[b]	0.0034	0.0051[b]	0.0034
	0.0004[c]	-0.0002	-0.408[c]	-0.2440	0.0052[c]	0.0033	0.0052[c]	0.0033
σ_{IS}	-0.0058	-0.0047	−−	−−	0.0083	0.0084	0.0060	0.0069
	-0.0010	-0.0069	-2.016	**-13.760**	0.0170	0.0168	0.0169	0.0153
	-0.0018	-0.0064	-0.876	-3.2200	0.0266	0.0241	0.0265	0.0232
$\beta_{Intercept}$	-0.0004	-0.0007	-0.0212	0.0352	0.0308	0.0287	0.0308	0.0286
	-0.0006	-0.0010	0.0288	0.0490	0.0423	0.0425	0.0423	0.0425
	-0.0014	-0.0050	0.0696	0.0246	0.0342	0.0369	0.0341	0.0369
β_{Slope}	0.0006	0.0003	0.1440	0.0670	0.0061	0.0045	0.0060	0.0045
	0.0010	-0.0090	0.3520	0.2300	0.0110	0.0105	0.0109	0.0105
	0.0014	0.0019	0.4170	0.4700	0.0189	0.0190	0.0189	0.0189
$\sigma^2_{Intercept}$	0.0038	0.0145	**19.020**	**7.2260**	0.0552	0.0357	0.0399	0.0326
	0.0214	0.0335	3.0543	4.7851	0.0792	0.0808	0.0762	0.0735
	0.0251	0.0206	**6.2860**	**5.1620**	0.0613	0.0529	0.0559	0.0487
σ^2_{slope}	0.0013	0.0016	**13.200**	**15.520**	0.0019	0.0027	0.0013	0.0022
	0.0046	0.0029	4.6320	2.936	0.0094	0.0060	0.0082	0.0052
	0.0162	0.0066	4.0440	1.662	0.0346	0.0187	0.0305	0.0175

Note: Same as Table 3a.

Table 8b: Data Condition 6 Using Initial Informative Priors – Parameter Evaluation Criteria Results

Parameter	Raw Bias		Relative Bias		Accuracy		Efficiency	
	FIRST	LAST	FIRST	LAST	FIRST	LAST	FIRST	LAST
	0.0006[a]	-0.0002	0.6032[a]	-0.1960	0.0059[a]	0.0031	0.0059[a]	0.0031
σ_e^2	0.0003[b]	-0.0005	0.2600[b]	-0.4920	0.0053[b]	0.0031	0.0053[b]	0.0031
	0.0007[c]	-0.0004	0.6960[c]	0.3800	0.0056[c]	0.0030	0.0055[c]	0.0030
	-0.0012	-0.0005	−−	−−	0.0086	0.0078	0.0086	0.0078
σ_{IS}	0.0007	-0.0017	1.3200	-3.3200	0.0181	0.0156	0.0181	0.0155
	0.0041	0.0010	2.0560	0.5020	0.0281	0.0238	0.0278	0.0238
	0.0031	0.0022	-0.1547	-0.1094	0.0361	0.0306	0.0362	0.0305
$\beta_{Intercept}$	0.0034	0.0025	-0.1680	-0.1234	0.0471	0.0430	0.0470	0.0430
	0.0027	0.0001	-0.1326	-0.0050	0.0374	0.0384	0.0373	0.0384
	-0.0002	0.0000	-0.0435	0.0070	0.0065	0.0047	0.0066	0.0047
β_{Slope}	0.0003	0.0002	0.0700	0.0420	0.0111	0.0105	0.0111	0.0105
	0.0008	0.0002	0.2080	0.0610	0.0190	0.0194	0.0190	0.0194
	0.0143	0.0042	**7.1478**	2.0960	0.0523	0.0375	0.0506	0.0373
$\sigma^2_{Intercept}$	0.0151	0.0188	2.1503	2.6880	0.0846	0.0719	0.0832	0.0694
	0.0069	0.0114	1.7160	2.8560	0.0691	0.0498	0.0688	0.0484
	0.0006	0.0002	**5.5466**	2.2000	0.0016	0.0020	0.0015	0.0020
σ^2_{Slope}	0.0030	0.0012	3.0160	1.1520	0.0091	0.0061	0.0086	0.0060
	0.0112	0.0035	2.8040	0.8810	0.0355	0.0192	0.0336	0.0189

Note: Same as Table 3a.

4 Discussion

Bayesian estimation operates by using prior information about the characteristics of parameters and the conditional likelihood of the data given the model parameters to arrive at a posterior distribution. The Bayesian Synthesis approach is based on this Bayesian estimation framework in which information obtained from one dataset serves to provide prior information for the analysis of the next dataset and this process continues sequentially until a single posterior distribution is created using all available datasets. While the benefits of using fused datasets have been repeatedly demonstrated in the literature (e.g., Curran & Hussong, 2009; Du et al., 2020; Hofer & Piccinin, 2009; Marcoulides, 2017b; Marcoulides & Grimm, 2017), and the estimates computed via a sequentially obtained final posterior distribution like those in the Bayesian Synthesis approach have also been shown to effectively aid in the accuracy of the estimation process (Du et al., 2020; Marcoulides, 2017b), what had not be determined was whether the order in which the data are sequentially analyzed has an impact on the obtained results.

The commonly accepted view in Bayesian estimation is that the order in which the data are analyzed should not be a concern due to the exchangeability assumption (de Finetti, 1972, 1974). Nevertheless, because Bayesian Synthesis utilizes point summary estimates of the posterior distributions instead of the full posterior distribution as required in standard Bayesian estimation, it possible that using point summary estimates of the posterior distributions may conceivably introduce some bias in the parameter estimates. Although past research has confirmed that the order of analysis does not meaningfully impact the final data fusion results obtained via Bayesian Synthesis (Marcoulides, 2017b), these conclusions were determined on the basis of analyzing datasets that were from similarly-sized and large samples. What was unresolved in the literature is whether exchangeability matters when the datasets being fused have substantially different sample sizes, as regularly occurs in empirical settings. Does beginning or ending the Bayesian Synthesis approach with the analysis of a large dataset produce a biased final posterior distribution when the other sequentially analyzed datasets are much smaller? This study examined via simulation the impact that the ordering of datasets might have on parameter estimates obtained when making use of the Bayesian Synthesis process in such data fusion design settings.

The results of the simulation study collectively highlighted the importance that the ordering of datasets can have on the estimation of growth model parameters when using the Bayesian Synthesis process. When the datasets being fused are of markedly different and much smaller sizes, ending the fusion and Bayesian estimation based on a large dataset produces a substantially biased (according to the measure of relative bias) final posterior distribution, particularly for the intercept and slope variance. Dissimilar longitudinal data design characteristics were also sometimes found to produce substantially biased final posterior distribution when implementing Bayesian Synthesis strategies. In longitudinal data design settings with fewer occasions of measurement and covering

varying ranges of ages, sizeable biased estimates were observed irrespective of whether a larger data set is fused first or last. Even in instances where there were numerous assessment occasions over a wider age range, the order of the fusing of the data sets still played a key role in the estimation, primarily with more sizable bias present when the large dataset was fused last. The results revealed that the order datasets of differing size are incorporated into the Bayesian Synthesis process along with the data design characteristics can impact the resulting parameter estimates and clearly calls into question the previously accepted notion of exchangeability of Bayesian estimation within the Bayesian Synthesis process.

Researchers planning on using the Bayesian Synthesis approach to data fusion should therefore be very careful how they elect to begin and end planned data fusion activities, especially in instances that involve the analyses of substantially large datasets among other much smaller datasets. Because Bayesian Synthesis uses point summary estimates from the analysis of one dataset as priors for the analysis of the next dataset, it is likely that this can introduce some bias in the Bayesian estimation. One explanation for this bias is that when the small datasets are being incorporated first, the informative priors that result from these small datasets are less reliable (contain sizable bias) and that this bias is then inevitably carried over to subsequent samples, resulting in a more biased final posterior distribution. Although it is commonly accepted that sample size can play an important role in the estimation of parameters, it is unclear in this context how much smaller the datasets can be relative to the other datasets. In this study, it was unmistakably determined that fusing a large dataset with smaller ones biased many of the parameter estimates provided by Bayesian Synthesis (particularly when measured by the relative bias criterion). But it is unclear what the ideal sample size needs to be in order to be used in the approach and ensure sufficiently stable parameter estimates. The current study fixed some of the data design characteristics in order to keep the scope of the work manageable. Given that a major benefit of Bayesian Synthesis is that data from multiple sources can be analyzed to obtain estimates of overall effects, examining other data size conditions under which this approach does not operate well is a natural extension to the current study. There is overall agreement among researchers that larger samples provide more stable estimates, but must all the fused datasets meet this requirement in order for Bayesian estimation exchangeability to hold? Although the current results indicated that exchangeability did not always hold in the examined data design scenarios involving growth curve models where there were differences in the size of the samples, the number of measurement occasions, time of first assessment and between assessments, as well as magnitude of the intercept and slope variances and covariances, it is of course possible that when modeling other statistical paradigms that different results may be observed. Empirical applications of the Bayesian Synthesis approach must make certain that the data fusion activities will provide researchers with unbiased parameter estimates.

Without doubt prior specification may be the largest advantage, yet potentially the greatest drawback, of implementing Bayesian methods in Bayesian

Synthesis. Priors enable researchers to include information from different data sources in a systematic manner. While it is recognized that imposing informative priors improves parameter estimates, especially with small sample sizes (Depaoli, 2014; Little, 2006), given that the true prior distribution is unknown in practice, researchers must be cautious about the impact that inaccurate priors have on parameter estimation in Bayesian Synthesis (Marcoulides, 2018). We also examined the impact of the order of incorporation in the Bayesian Synthesis process using initial informative data-dependent priors to analyze the first dataset in the data fusion process. Across the various design conditions, moderate and substantial bias was primarily found when the large dataset was analyzed last. These results are consistent with those found when using initial non-informative or diffuse priors to analyze the first dataset in the data fusion process. Future research studies should therefore expand further on our findings and examine additional data design conditions and settings.

The process of sequentially updating information to arrive at conclusions undeniably has a substantiated place in data analyses and Bayesian Synthesis can play a key role in helping researchers address questions not always achievable with a single study. Although additional research needs to be done regarding when Bayesian Synthesis is most useful and when it might prove to be problematic, the foundations for the continued use of this data fusion process are evident. We caution researchers to remain mindful of the limitations identified in this study when integrating data from different sources.

References

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using mplus : Technical implementation..

Bandalos, D. L., & Gagne, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92–110). New York, NY: Guildford Press.

Bandalos, D. L., & Leite, W. L. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course (2nd ed)* (pp. 564–666). Greenwich, CT.

Bhattacharya, B., & Saha, B. (2015). Community model: A new data fusion filter paradigm. *American Journal of Advanced Computing*, *2*, 25–31. doi: https://doi.org/10.15864/ajac.1303

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple datasets. *Psychological Methods*, *14*, 81–100. doi: https://doi.org/10.1037/a0015914

de Finetti, B. (1972). *Probability, induction, and statistics*. New York, NY: John Wiley & Sons.

de Finetti, B. (1974). *Theory of probability (Vol. I and Vol. II)*. New York, NY: John Wiley & Sons.

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling*, *21*, 239–252. doi: https://doi.org/10.1080/10705511.2014.882686

Du, H., Bradbury, T. N., Lavner, J. A., Meltzer, A. L., McNulty, J. K., Neff, L. A., & Karney, B. R. (2020). A comparison of Bayesian synthesis approaches for studies comparing two means: A tutorial. *Research Synthesis Methods*, *11*, 36–65. doi: https://doi.org/10.1002/jrsm.1365

Finch, W. H., & Miller, J. E. (2019). The use of incorrect informative priors in the estimation of MIMIC model parameters with small sample sizes. *Structural Equation Modeling*, *26*, 497–508. doi: https://doi.org/10.1080/10705511.2018.1553111

Fujimoto, K. A., Gordon, R. A., Peng, F., & Hofer, K. G. (2018). Examining the category functioning of the ECERS-R across eight data sets. *AERA Open*, *4*, 1–16. doi: https://doi.org/10.1177/2332858418758299

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd Edition)*. CRC.

Hallquist, M., & Wiley, J. (2014). *Mplusautomation: Automating Mplus model estimation and interpretation* (Vol. 06-3.).

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, *14*, 150–164. doi: https://doi.org/10.1037/a0015566

Johnson, R. M., & Guttmannova, K. (2019, jan). Marijuana use among adolescents and emerging adults in the midst of policy change: Introduction to the special issue. *Prevention Science*, *20*(2), 179–184. doi: https://doi.org/10.1007/s11121-019-0989-7

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, *60*, 213–223. doi: https://doi.org/10.1198/000313006x117837

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling*, *23*, 354–367. doi: https://doi.org/10.1080/10705511.2015.1057285

Marcoulides, K. M. (2017a). *A Bayesian synthesis approach to data fusion using augmented data-dependent priors (Doctoral dissertation)* (Unpublished doctoral dissertation). (Retrieved from ProQuest Dissertations & Theses Global Database. (Accession No. 10276332).)

Marcoulides, K. M. (2017b). A Bayesian synthesis approach to data fusion using data-dependent priors. *Multivariate Behavioral Research*, *52*(1), 111–112. doi: https://doi.org/10.1080/00273171.2016.1263927

Marcoulides, K. M. (2018). Careful with those priors: A note on Bayesian estimation in two-parameter logistic item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *16*(2), 92–99. doi: https://doi.org/10.1080/15366367.2018.1437305

Marcoulides, K. M., & Grimm, K. J. (2017). Data integration approaches to lon-

gitudinal growth modeling. *Educational and Psychological Measurement*, *77*, 971–989. doi: https://doi.org/10.1177/0013164416664117

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. doi: https://doi.org/10.1037/a0039400

McNeish, D. (2016, jun). On using bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. doi: https://doi.org/10.1080/10705511.2016.1186549

Miocevic, M., Levy, R., & Savord, A. (2020). The role of exhangeability in sequential updating of findings from small studies and the challenges of identifying exchangeable data sets. In . M. M. R. van de Schoot (Ed.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 13–29). London, UK: Routledge.

Muthen, L. K., & Muthen, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620. doi: https://doi.org/10.1207/s15328007sem0904_8

Muthen, L. K., & Muthen, B. O. (2017). *Mplus user's guide (8th edition)*. Los Angeles, CA: Muthen & Muthen.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001, apr). Monte carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. doi: https://doi.org/10.1207/s15328007sem0802_7

Preston, K. S. J., Gottfried, A. W., Park, J. J., Manapat, P. D., Gottfried, A. E., & Oliver, P. H. (2018). Simultaneous linking of cross-informant and longitudinal data involving positive family relationships. *Educational and Psychological Measurement*, *78*, 409–429. doi: https://doi.org/10.1177/0013164417690198

R Development Core Team. (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saris, W., & Satorra, A. (2018). The pooled data approach for the estimation of split-ballot multitrait-multimethod experiments. *Structural Equation Modeling*, *25*, 659–672. doi: https://doi.org/10.1080/10705511.2018.1431543

Shi, D., & Tong, X. (2017). The impact of prior information on Bayesian latent basis growth model estimation. *SAGE Open*, *7*. doi: https://doi.org/10.1177/2158244017727039

Shi, D., & Tong, X. (2018). Bayesian robust two-stage estimation with nonnormal missing data. *Multivariate Behavioral Research*, *53*, 127–127. doi: https://doi.org/10.1080/00273171.2017.1404894

Journal of Behavioral Data Science, 2022, *2* (1), 106–123.
DOI: https://doi.org/10.35566/jbds/v2n1/p4

The Role of Personality in Trust in Public Policy Automation

Philip D. Waggoner*[1] and Ryan Kennedy[2]

[1] Columbia University ,YouGov America
pdw2119@columbia.edu
[2] University of Houston

Abstract. Algorithms play an increasingly important role in public policy decision-making. Despite this consequential role, little effort has been made to evaluate the extent to which people trust algorithms in decision-making, much less the personality characteristics associated with higher levels of trust. Such evaluations inform the widespread adoption and efficacy of algorithms in public policy decision-making. We explore the role of major personality inventories – need for cognition, need to evaluate, the "Big 5" – in shaping an individual's trust in public policy algorithms, specifically dealing with criminal justice sentencing. To explore personality in this context, we fielded an original survey experiment aimed at assessing the impact of varying advice sources on forecasting criminal recidivism, conditioned by personality traits. We found strong correlations between all personality types and general levels of trust in automation, as expected. Further, we uncovered evidence that need for cognition increases the weight given to advice from an algorithm relative to humans, and "agreeableness" decreases the distance between respondents' expectations and advice from a judge, relative to advice from a crowd.

Keywords: Personality · Trust in automation · Public policy · Decision-making

1 Introduction

Algorithms are increasingly important in public policy implementation (Kennedy, Waggoner, & Ward, 2022). Algorithms assist officials in major US cities to allocate resources (O'Brien, 2015), judges in detecting gerrymandering (Bernstein & Duchin, 2017), and the military to control weapons (Scharre, 2018). Recently, algorithms have also begun to play a role in criminal sentencing, where algorithms are used by judges to inform expectations on a defendant's probability of recidivating (Waggoner & Macmillen, 2021). Such a hybrid-decision making process between humans and algorithms influences the parameters, duration, and severity of sentencing (Dressel & Farid, 2018).

Despite the rise in interest about automation and algorithms, little attention has been paid in public policy to algorithms or the psychological factors that influence trust in them. Horowitz (2016) explored situations under which people approve development of autonomous weapons systems, but this reveals little about the underlying trust people have in algorithms in practice. Further, some have debated whether individuals place low levels of trust in algorithms, "algorithm aversion" (Dietvorst, Simmons, & Massey, 2015), or high levels of trust, "algorithm bias" (Logg, 2016). Still, very little attention has been paid to how individuals' psychological characteristics might influence attitudes towards algorithms. Instead, the literature tends to focus on demographic or cultural factors (Hoff & Bashir, 2015).

To address this gap, we recently fielded a criminal sentencing survey experiment and leveraged three major inventories of psychological measures of personality to explore who is more or less trusting of algorithms: "need for cognition" (NC) (Cacioppo & Petty, 1982), "need to evaluate" (NE) (Bizer et al., 2004), and the "Big 5" (Norman, 1963).

The survey experiment was primarily interested in assessing the impact of varying advice sources (judge, algorithm, a "crowd" of peers) on respondents' forecasts of criminal recidivism. Of primary interest was the conditioning role of personality in this forecasting effort. The details of and findings from the experiment are detailed throughout the remainder of the paper.

1.1 Personality Inventories

The first inventory, need for cognition (NC), is associated with individuals who have a strong desire to learn and grow (Cacioppo & Petty, 1982). Some previous studies on NC in similar contexts have suggested that when high NC individuals are asked to undertake a task in which they are given little information and then provided expert advice, they are more likely to assign greater weight to that advice, rather than relying on heuristics (Leippe, Eisenstadt, Rauch, & Seib, 2004). This suggests that high NC individuals will be more likely to *take* advice insofar as they view that advice as "expert," given their more elaborate processing of information (Sicilia, Ruiz, & Munuera, 2005).

Our second, need to evaluate (NE), is associated with individuals who tend to generate and retain their own attitudes (Bizer et al., 2004). This "self-monitoring" personality is also associated with the need to control (Snyder, 1974) and constantly evaluate social surroundings (Jarvis & Petty, 1996). Such attributes induce greater reliance on intuition over outside sources. Past work has demonstrated that high NE individuals tend to make spontaneous judgements in *response* to stimuli (Tormala & Petty, 2001). This "on-line" form of information processing suggests that when people come into contact with outside information, their personality plays a key role in determining their levels of acceptance of the information. Resultant attitudes are much stronger than those in the alternative, "memory-based" processing (Bizer, Tormala, Rucker, & Petty, 2006). Other studies have also leveraged NE to explain information processing (Druckman & Nelson, 2003). For our context, we expect high NE individuals exhibit

greater reliance on their own intuition compared to other sources, which should lead to distrust. When a high NE individual is confronted by advice from an outside source, we expect these individuals to be less trusting of advice, *regardless* of its origin. This is in line with recent work suggesting that errors experienced from an algorithm provoke a stronger distrust of that advice than do errors experienced from other sources (Dietvorst et al., 2015).

For measurement, we used the two-item battery for each personality type (NC and NE), totaling four questions for both personality types. Question wording is in the Appendix. Of note, though there is a tradeoff between "internal reliability and brevity" (Gerber, Huber, Doherty, & Dowling, 2011), we opted for the smaller battery for two reasons. First, it is the exact same approach as using the common, reliable TIPI inventory to measure each of the Big 5 traits (Gosling, Rentfrow, & Swann Jr, 2003). Both approaches uses two items per personality trait to generate a measure. Second, we wanted to ensure high response rates, given the inclusion of the personality batteries in addition to our main experiment. To minimize the burden on the respondent and with the "the benefit of being short enough to be included in large political surveys," (Gerber et al., 2011, 268), we opted for the smaller battery. Ultimately, we selected these measures of personality given their widespread use in a variety of fields including political science (Jost, Glaser, Kruglanski, & Sulloway, 2003), public policy (Sargent, 2004), psychology (Cacioppo, Petty, & Feng Kao, 1984), and others (Luttrell, Petty, & Xu, 2017).

Turning now to the Big 5, we use only the "agreeableness" and "openness to experience" traits in our study as they can be most clearly linked to trust in automation. We selected only two instead of all five traits, because, as Gerber et al. (2011) note, "in most cases only some of the Big Five traits significantly predict outcomes of interest" (268). Our approach is similar to other studies on the role of the Big 5 in behavior that select only the specific personality traits that can be clearly linked to substantive phenomena (Quintelier, 2014).

For agreeableness, Gerber et al. (2011) and John and Srivastava (1999) note that "agreeableness contrasts a prosocial and communal orientation toward others with antagonism and includes traits such as altruism, tender-mindedness, trust, and modesty." Agreeableness is also associated with social conformity (Fiske, 1949) and compliance (Digman & Takemoto-Chock, 1981). In our context, being given advice from an "expert," and then asked whether they wish to update their expectation, we expect agreeable individuals should positively respond to the advice-giver, regardless of the source of advice. In an effort to conform to the reigning wisdom via the advice treatment, individuals who are high on agreeableness should trust automation, positively weight expert advice, and also align with the advice-giver.

Second, openness is associated with originality (Gerber et al., 2011; John & Srivastava, 1999), intellectual curiosity (Peabody & Goldberg, 1989), and an eagerness to learn (Barrick & Mount, 1991). As individuals who are open to experiences come into contact with outside advice in an unfamiliar realm, they

should positively respond to the advice treatment across all three measures of trust discussed below.

We leveraged the Ten-Item Personality Inventory (TIPI) (Gosling et al., 2003) to measure these traits. Two items containing personality adjectives are associated with each trait, with one phrase coded normally and the other reverse coded (e.g., for "agreeableness": *item 7* = sympathetic, warm and *item 2* (reverse coded) = critical, quarrelsome).

2 Method

2.1 Participants

We utilized Amazon's Mechanical Turk (*MTurk*) to recruit 395 subjects, each of whom were paid $2.00 for participation. MTurk is a valid, widely used platform to field similar political, psychological, and social experiments such as ours (Clifford, Jewell, & Waggoner, 2015). Additional details of the study design are included in the Appendix.

2.2 Procedure

Our study contains observational (general trust) and experimental (behavioral impact) components. For the observational component, respondents were given an eight-item battery of questions related to degrees of trust in automation (Kim, Ferrin, & Rao, 2008). Respondents were asked their level of agreement on a scale from 1 (strongly disagree) to 7 (strongly agree) for statements like, "Using algorithms improves the output quality for organizations." These were aggregated into a 7 point scale where 7 indicates high trust in algorithms, while 1 indicates low trust. The wording for all of the items is available in the supporting information. This scale is the dependent variable for the first stage of the analysis, which is analyzed using OLS regression and presented in Table 1.

For the experimental component, respondents were asked to forecast the probability of a defendant committing *another* crime within two years for one of eight real, randomly selected criminal profiles based on criminal history and defendant demographic characteristics. Then, the respondent was given "advice" from a source (listed below), and asked whether they wanted to update predictions or leave them the same (manual entry required both times). The shifts in respondents' predictions (or lack thereof) is the quantity of interest in our study. We included two attention checks throughout to minimize satisficing (Hauser & Schwarz, 2016). Specifically, respondents were warned if they missed one attention check, and then were removed and not paid if they failed both. About 80% of respondents who attempted the survey passed the checks and completed the survey.

The presentation of our criminal profiles mimics the formatting of Dressel and Farid (2018), which was shown to be a sufficient amount of detail for an average MTurk participant to make an informed judgment, with expected accuracy

similar to the popular "COMPAS" algorithm. The full wording is available in the supporting information. We randomly selected 20 pre-trial defendants from the 2013-14 from Broward County, FL database, who all had a risk scores between 2 and 8 (derived from the COMPAS algorithm, which ranked defendants from 1 to 10, with 10 being the most likely to recidivate). This pool of defendants was winnowed when the crime involved was obscure, and then reduced again randomly to reduce the task burden on respondents, which left us with eight profiles.

For each profile, respondents were randomly assigned to one of the three advice conditions: judge with 10 years of experience in criminal sentencing; computer algorithm designed by computer scientists and criminal justice officials; average of a previous survey of 300 Turkers. The treatment conditions are coded as separate dummy variables for whether the individual saw advice from an algorithm or a judge in the scenario, with the previous MTurk survey as the baseline condition. And, in addition to the main personality predictors, we control for several common factors in public policy experiments, including age, education, gender, and partisanship.

We evaluate two measures of trust. The first measure is "weight of advice" (Gino & Moore, 2007; Logg, 2016). This variable is calculated as $\mid u_{2i} - u_{1i} \mid / \mid a_i - u_{1i} \mid$, where u_{2i} is respondent i's final assigned probability for recidivism, u_{1i} is their initial prediction, and a_i is the advice they were given from one of the sources. A score of 1 suggests the respondent only used the advice from the source, where as 0.5 suggests they weighted the source and their prediction equally, and 0 means the respondent ignored the advice. Our second measure is the average distance to advice, measured as $\mid a_i - u_{ri} \mid$. Lower values indicate that there was less distance between the respondent's final forecast and the advice they were given.

We modeled the weight and distance measures by fitting multilevel regressions to the data after pooling across all criminal profiles and specifying varying intercepts for defendant descriptions and respondent. Multilevel models were chosen to account for unobserved heterogeneity on both the individual respondent and scenario level. This provides an efficient and accurate estimates for experiments where respondents evaluate multiple, different scenarios (Gelman & Hill, 2006). The model was specified as

$$y_{ijk} = a_{ijk} + \zeta_j + \phi_k + \beta * X + e_{ijk} \tag{1}$$

where a_{ij} is the overall intercept, $\zeta_j \sim N(0,1)$ is the random intercept based on the defendant description, $\phi_k \sim N(0,1)$ is the random intercept based on the individual respondent, β is an array of coefficients for the treatments X, and e_{ij} is the error term. Results are presented in Table 2.

3 Results

For the observational analysis, note the significance and large magnitudes of effects for all personality indicators in the top four rows of Table 1.[1] High NE respondents are less likely to trust algorithms ($\beta = -0.14$), compared to those higher on NC ($\beta = 0.25$), agreeableness ($\beta = 0.04$), and openness ($\beta = 0.07$), all of whom are eager to learn. The latter group of respondents is more trusting of automation in line with expectations.

Table 1. The Impact of Personality on Trust in Automation

	Dependent variable:	
	Trust in Automation	
	(1)	(2)
Need for Cognition	0.245^{***} (0.023)	
Need to Evaluate	-0.136^{***} (0.021)	
Agreeableness		0.040^{**} (0.016)
Openness to Experience		0.066^{***} (0.015)
Age	0.005^{***} (0.002)	0.007^{***} (0.002)
Education	0.131^{***} (0.029)	-0.020 (0.040)
Female	-0.212^{***} (0.035)	-0.336^{***} (0.040)
Partisanship	-0.047^{***} (0.009)	-0.028^{***} (0.010)
Algorithm Condition	-0.000 (0.00000)	-0.000 (0.00000)
Judge Condition	-0.000 (0.00000)	0.000 (0.00000)
Constant	3.783^{***} (0.124)	4.047^{***} (0.173)
N	3,022	2,233
Log Likelihood	32,311.620	24,075.100
Akaike Inf. Crit.	$-64{,}599.240$	$-48{,}126.190$
Bayesian Inf. Crit.	$-64{,}527.070$	$-48{,}057.660$

Note: *p<0.1; **p<0.05; ***p<0.01

The experimental stage exploring the impact of personality on behavioral tasks seen in Table 2 and Figure 1. Here our N is higher because each respondent evaluated 8 defendant profiles. NC plays a strong conditioning role in the relative weight respondents' assign to advice across both "expert" conditions in comparison to the baseline category. The degree to which NC conditions trust in *algorithms* is nearly doubled that of the *judge* condition ($\beta = 0.09$ compared to $\beta = 0.05$). Further, the weight effect is opposite for NE individuals in the algorithm condition ($\beta = -0.04$), and indistinguishable from zero in the

[1] Of note, the trust in automation index is measured at the individual-level, not the scenario-level. Hence the larger N in the tables, relative to the number of individual recruited subjects.

judge condition. Results are similar for high openness personality types in their weighting of algorithmic advice relative to humans.

Table 2. The Impact of Personality on Behavior

	Dependent variable:			
	Weight	Distance	Weight	Distance
	(1)	(2)	(3)	(4)
Need for Cognition	−0.047*** (0.018)	1.687* (0.892)		
Need to Evaluate	0.012 (0.017)	−0.964 (0.839)		
Agreeableness			0.018 (0.013)	−0.501 (0.642)
Openness			−0.017 (0.012)	−0.126 (0.606)
Age	0.0002 (0.001)	0.008 (0.043)	0.001 (0.001)	−0.012 (0.051)
Education	0.010 (0.019)	−1.879** (0.857)	0.010 (0.022)	−0.571 (1.008)
Female	0.015 (0.020)	−0.678 (0.930)	−0.001 (0.025)	0.758 (1.111)
Partisanship	0.004 (0.005)	−0.187 (0.231)	0.005 (0.006)	−0.128 (0.279)
Algorithm Cond.	−0.072 (0.079)	−1.701 (4.101)	0.120 (0.097)	−9.706* (5.081)
Judge Cond.	0.006 (0.083)	−1.969 (4.307)	−0.030 (0.097)	1.395 (5.144)
Alg. x NC	0.093*** (0.024)	−3.007** (1.251)		
Alg. x NE	−0.035* (0.021)	1.905* (1.106)		
Judge x NC	0.054** (0.022)	−1.929* (1.168)		
Judge x NE	−0.033 (0.022)	1.645 (1.140)		
Alg. x Agreeable			−0.024 (0.016)	0.610 (0.851)
Alg. x Openness			0.028* (0.015)	0.228 (0.801)
Judge x Agreeable			0.031* (0.016)	−2.216** (0.876)
Judge x Openness			−0.005 (0.016)	1.275 (0.854)
Constant	0.225** (0.093)	30.664*** (6.182)	0.076 (0.115)	32.606*** (7.002)
N	3,022	3,022	2,233	2,233
Log Likelihood	−403.901	−12,702.400	−349.032	−9,388.203
Akaike Inf. Crit.	839.802	25,436.800	730.063	18,808.410
Bayesian Inf. Crit.	936.021	25,533.020	821.441	18,899.780

*p<0.1; **p<0.05; ***p<0.01

Notably, NC strongly conditions trust in algorithms, but less so compared to advice from crowds or human experts. This is seen most clearly when comparing panels (a) and (b) in Figure 1. The gaps between the fit lines are more distinct in the algorithm condition (a) compared to the judge condition (b). There is only a modest distinction at the tails in the judge condition.

Across all treatment conditions, the advice given by all three sources was the same, and we found no differences when we presented values centered around those derived from the COMPAS algorithm or when the advice was randomly chosen.

Of note, for the weight and distance multilevel models, to test if there was any impact of varying the treatments by scenario or by respondent, we ran-

domly allocated half of our respondents to each type of assignment. We found no difference (i.e., no detection of study purpose) in the results.

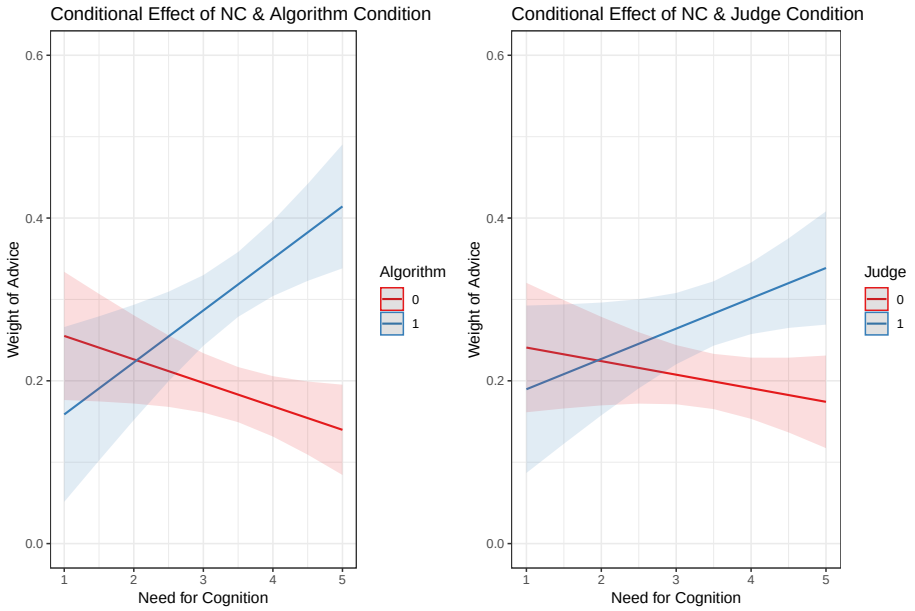


Figure 1. Conditional Impacts of NfC on Behavior

4 Discussion

Overall, we found that personality influences trust in automation, as well as behavioral tasks related to public policy decision-making. In the first stage, there was a pronounced impact of personality on general levels of trust. In line with research finding high levels of trust in algorithms (Goddard, Roudsari, & Wyatt, 2011), the significant conditioning role of these personality inventories suggests that personalities associated with intellectual curiosity, agreeableness, openness to advice-givers, as well as being highly aware of environments and more skeptical are strongly associated with levels of trust in automation. The former group comprised of individuals who are more accepting of new information and experiences is more trusting, while the latter group, who tends to be threatened by exogenous sources of information, is less trusting.

Regarding changes in respondents' *behavioral* indicators of trust, high NC individuals are much more trusting of algorithms than of the wisdom of the crowd or, to a lesser extent, a human expert. And the NE personality trait, which we expected to be threatened by exogenous advice, weighs advice from algorithms

less than human advice sources. Surprisingly, no effects were observed for agreeable individuals in the algorithmic advice condition for weighting, though there were weak effects for judges. Strikingly, high NE and NC individuals reacted to algorithmic advice over all other advice sources, though the effect size is nearly doubled for NC individuals compared to NE individuals and is more statistically stable. We also saw a significant effect of personality conditioning behavior in decision making tasks, especially related to their trust in advice from an algorithm.

Though a blend of significant and null results, we remain encouraged by our findings for two reasons. First, we uncovered strong evidence of personality influencing behavior and general levels of trust in automation, in line with our main goal. Given the newness of this topic, these results are useful for motivating future work on trust in automation and personality. Second, in line with Gerber et al. (2011), it would be unrealistic to expect *all* personality measures to explain *all* behavior. Of the Big 5 they note, "these traits have predictive power in an impressive variety of domains but are not universal predictors of all outcomes" (268). Our results corroborate this sentiment that personality plays a role in trust in automation, though it does not explain the breadth of general trust *and* behavior.

Regarding generalizability, while people generally trust algorithmic advice relative to other advice sources, levels of trust are influenced by personality traits. As not all people retain the same personalities, not all people equally trust algorithms to make consequential decisions.

5 Limitations and Future Directions

While we offer a starting place for future work on personality and trust in automation, a key limitation of our study is focusing only on criminal justice. Should we expect similar results in other subfields, such as automation in medicine, for example? Also, though Dietvorst et al. (2015) demonstrate trust in algorithms wanes when mistakes are introduced, this phenomenon may be more likely for high NE individuals relative to high NC individuals, given the starting place of skepticism for high NE individuals. Further, algorithm aversion may not be detectable for high NC individuals, while it may drive levels of trust for high NE individuals. Or, do the other three Big 5 personality traits (extroversion, conscientiousness, and emotional stability) impact trust in automation? In sum, we suggest researchers in this realm consider personalities to provide a fuller picture of trust in a variety of subfields.

An additional limitation that may be addressed in future work is the nature of MTurk respondents in general, in that they are typically higher educated and more liberal for example (Berinsky, Huber, & Lenz, 2012; Clifford et al., 2015), and thus may be more likely to trust automation. Such a possibility suggests the potential for future and different samples to yield potentially different results. More analysis and experiments in this vein would deepen the impact of our initial findings in this research.

6 Concluding Remarks

In this study, we have demonstrated that personality plays a strong role in impacting individuals' levels of trust in automation as they make public policy decisions. We bring psychology into the trust in automation discussion for several reasons. First, such an approach offers a baseline for understanding the role of innate, heritable characteristics and their influence on trust in automation.[2] Such an understanding makes it clearer where to look for greater or lesser trust in algorithms, and where the basis of trust lies. These psychological characteristics are also widely used in many fields to describe human behavior both inside (Sargent, 2004) and outside (Hill, Foster, Sofko, Elliott, & Shelton, 2016) of public policy. Given the rapid increase of algorithms and algorithmic advice in everyday life (Logg, 2016), the role of psychological characteristics conditioning virtually all human behavior (Eysenck, 1963), and also the recent surge in research on algorithmic transparency (Rudin & Ustun, 2018), our study offers a timely exploration of the intersection of trust in automation and personality.

Acknowledgement

The authors thank Scott Clifford for his comments on a previous version of this paper. All errors are the authors'. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17061500008. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, *44*(1), 1–26. doi: https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political analysis*, *20*(3), 351–368. doi: https://doi.org/10.1093/pan/mpr057

Bernstein, M., & Duchin, M. (2017). A formula goes to court: Partisan gerrymandering and the efficiency gap. *Notices of the AMS*, *64*(9). doi: https://doi.org/10.1090/noti1573

[2] In using the terms "innate" and "heritable," we are referring to the vast work on the stability of personality throughout one's life (Caspi, Roberts, & Shiner, 2005), the heritable nature of personality (Van Gestel & Van Broeckhoven, 2003), the genetic aspects of personality (Canli, 2008), and even the biological link with personality (DeYoung et al., 2010).

Bizer, G. Y., Krosnick, J. A., Holbrook, A. L., Christian Wheeler, S., Rucker, D. D., & Petty, R. E. (2004). The impact of personality on cognitive, behavioral, and affective political processes: The effects of need to evaluate. *Journal of personality*, *72*(5), 995–1028. doi: https://doi.org/10.1111/j.0022-3506.2004.00288.x

Bizer, G. Y., Tormala, Z. L., Rucker, D. D., & Petty, R. E. (2006). Memory-based versus on-line processing: Implications for attitude strength. *Journal of Experimental Social Psychology*, *42*(5), 646–653. doi: https://doi.org/10.1016/j.jesp.2005.09.002

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116. doi: https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of personality assessment*, *48*(3), 306–307. doi: https://doi.org/10.1207/s15327752jpa4803_13

Canli, T. (2008). Toward a molecular psychology of personality. *Handbook of personality: Theory and research*, 311–327.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annu. Rev. Psychol.*, *56*, 453–484. doi: https://doi.org/10.1146/annurev.psych.55.090902.141913

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from mechanical turk valid for research on political ideology? *Research & Politics*, *2*(4), 2053168015622072. doi: https://doi.org/10.1177/2053168015622072

DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological science*, *21*(6), 820–828. doi: https://doi.org/10.1177/0956797610370159

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114. doi: https://doi.org/10.1037/xge0000033

Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. *Multivariate behavioral research*, *16*(2), 149–170. doi: https://doi.org/10.1207/s15327906mbr1602_2

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), 55–80. doi: https://doi.org/10.1126/sciadv.aao5580

Druckman, J. N., & Nelson, K. R. (2003). Framing and deliberation: How citizens' conversations limit elite influence. *American Journal of Political Science*, *47*(4), 729–745. doi: https://doi.org/10.1111/1540-5907.00051

Eysenck, H. (1963). *The biological basis of personality*. Routledge. doi: https://doi.org/10.4324/9781351305280

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings

from different sources. *The Journal of Abnormal and Social Psychology*, *44*(3), 329. doi: https://doi.org/10.1037/h0057198

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press. doi: https://doi.org/10.1017/cbo9780511790942

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The big five personality traits in the political arena. *Annual Review of Political Science*, *14*. doi: https://doi.org/10.1146/annurev-polisci-051010-111659

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, *20*(1), 21–35.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. doi: https://doi.org/10.1136/amiajnl-2011-000089

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528. doi: https://doi.org/10.1016/s0092-6566(03)00046-1

Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, *48*(1), 400–407.

Hill, B. D., Foster, J. D., Sofko, C., Elliott, E. M., & Shelton, J. T. (2016). The interaction of ability and motivation: Average working memory is required for need for cognition to positively benefit intelligence and the effect increases with ability. *Personality and Individual Differences*, *98*, 225–228. doi: https://doi.org/10.1016/j.paid.2016.04.043

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Horowitz, M. C. (2016). Public opinion and the politics of the killer robots debate. *Research & Politics*, *3*(1), 2053168015627183.

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of personality and social psychology*, *70*(1), 172. doi: https://doi.org/10.1037/0022-3514.70.1.172

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*(1999), 102–138.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological bulletin*, *129*(3), 339. doi: https://doi.org/10.4324/9781315175867-5

Kennedy, R. P., Waggoner, P. D., & Ward, M. M. (2022). Trust in public policy algorithms. *The Journal of Politics*, *84*(2). doi: https://doi.org/10.1086/716283

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision support systems*, *44*(2), 544–564. doi: https://doi.org/10.1016/j.dss.2007.07.001

Leippe, M. R., Eisenstadt, D., Rauch, S. M., & Seib, H. M. (2004). Timing of eyewitness expert testimony, jurors' need for cognition, and case strength as determinants of trial verdicts. *Journal of Applied Psychology*, *89*(3), 524. doi: https://doi.org/10.1037/0021-9010.89.3.524

Logg, J. M. (2016). *When do people rely on algorithms?* (Unpublished doctoral dissertation). University of California, Berkeley.

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178–183. doi: https://doi.org/10.1016/j.jesp.2016.09.006

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574. doi: https://doi.org/10.1037/h0040291

O'Brien, D. T. (2015). Custodians and custodianship in urban neighborhoods: A methodology using reports of public issues received by a city's 311 hotline. *Environment and Behavior*, *47*(3), 304–327.

Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of personality and social psychology*, *57*(3), 552. doi: https://doi.org/10.1037/0022-3514.57.3.552

Quintelier, E. (2014). The influence of the big 5 personality traits on young people's political consumer behavior. *Young Consumers*, *15*(4), 342–352. doi: https://doi.org/10.1108/yc-09-2013-00395

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *48*(5), 449–466. doi: https://doi.org/10.1287/inte.2018.0957

Sargent, M. J. (2004). Less thought, more punishment: Need for cognition predicts support for punitive responses to crime. *Personality and Social Psychology Bulletin*, *30*(11), 1485–1493. doi: https://doi.org/10.1177/0146167204264481

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. New York: WW Norton and Company.

Sicilia, M., Ruiz, S., & Munuera, J. L. (2005). Effects of interactivity in a web site: The moderating effect of need for cognition. *Journal of advertising*, *34*(3), 31–44. doi: https://doi.org/10.1080/00913367.2005.10639202

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of personality and social psychology*, *30*(4), 526. doi: https://doi.org/10.1037/h0037039

Tormala, Z. L., & Petty, R. E. (2001). On-line versus memory-based processing: The role of "need to evaluate" in person perception. *Personality and social psychology bulletin*, *27*(12), 1599–1612. doi: https://doi.org/10.1177/01461672012712004

Van Gestel, S., & Van Broeckhoven, C. (2003). Genetics of personality: are we making progress? *Molecular psychiatry*, *8*(10), 840. doi: https://doi.org/10.1038/sj.mp.4001367

Waggoner, P. D., & Macmillen, A. (2021). Pursuing open-source de-

velopment of predictive algorithms: the case of criminal sentencing algorithms. *Journal of Computational Social Science*, 1–21. doi: https://doi.org/10.1007/s42001-021-00122-y

Appendix

A Support Information

A.1 Task Wording

It has generally been found that even untrained individuals can do very well, sometimes even better than trained people or computer algorithms, at determining the likelihood of a person committing another crime after their initial arrest.

We are interested in knowing whether this accuracy can be further improved by combining individual judgement with the advice of crowds, experts, or algorithms. In what follows, you will be given an actual arrest record for a person arrested in Broward County, Florida. We already know whether the person committed another crime within the next two years. You will be asked to give us a probability of the person re-offending along the following lines.

We have collected advice from several sources:

- Several Mechanical Turk surveys of people like yourself.
- A judge with over 10 years of experience.
- A machine learning algorithms, developed by computer scientists and criminal justice experts, that use historic recidivism data to predict probability of re-offending.

Warning: There are attention checks in this survey. We reserve the right to deny payment if a participant fails these checks, as that indicates the participant is not actually doing the tasks.

A.2 Defendant Profiles

The defendant is a male aged 22. They have been charged with: Possession of Cocaine. This crime is classified as a felony. They have been convicted of 0 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

The defendant is a male aged 38. They have been charged with: Manufacturing Cannabis/Marijuana. This crime is classified as a felony. They have been convicted of 3 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

The defendant is a male aged 23. They have been charged with: Grand Theft. This crime is classified as a felony. They have been convicted of 3 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

The defendant is a male aged 27. They have been charged with: Possession of Meth. This crime is classified as a felony. They have been convicted of 5 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

The defendant is a male aged 24. They have been charged with: Driving with a Revoked License. This crime is classified as a felony. They have been convicted of 2 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

The defendant is a female aged 33. They have been charged with: Child Neglect. This crime is classified as a felony. They have been convicted of 1 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

The defendant is a male aged 22. They have been charged with: Disorderly Conduct. This crime is classified as a misdemeanor. They have been convicted of 0 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

The defendant is a male aged 24. They have been charged with: Resisting an Officer with Violence. This crime is classified as a felony. They have been convicted of 0 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

A.3 Examples of Treatment

A group of 200 people recruited from Mechanical Turk, on average rated the defendant as 80% likely to commit another felony crime within the next two years.

Previously, you forecast that the defendant was [RESPONDENT'S PREVIOUS FORECAST] likely to commit another felony crime within the next two years.

If you would like to update your forecast, you can do so now. If not, just enter the same numbers as you entered previously.

A judge with more than 10 years of experience rated the defendant as 80% likely to commit another felony crime within the next two years.

Previously, you forecast that the defendant was [RESPONDENT'S PREVIOUS FORECAST] likely to commit another felony crime within the next two years.

If you would like to update your forecast, you can do so now. If not, just enter the same numbers as you entered previously.

An algorithm developed by computer scientists and criminal justice researchers, based on a statistical analysis of thousands of past defendant records, rated the defendant as 80% likely to commit another felony crime within the next two years.

Previously, you forecast that the defendant was [RESPONDENT'S PREVIOUS FORECAST] likely to commit another felony crime within the next two years.

If you would like to update your forecast, you can do so now. If not, just enter the same numbers as you entered previously.

A.4 MTurk Study Specifics

In addition to the specifics of the design included in the manuscript, below are some additional specific items related to fielding the study on MTurk:

1. **Approval Rate**: HIT Approval Rate > 95%
2. **Location**: United States
3. **Study Description**: Respondents will be asked to evaluate a series of real criminal profiles and asked to predict the likelihood of recidivism with and without the help of advice.
4. **Keywords**: survey, criminal justice, forecasting, predication

A.5 Personality Inventories

A.5.1 NC and NE

Please indicate the extent to which these statements are characteristic or uncharacteristic of you (On a scale from 1 to 5, with being extremely characteristic and 5 being extremely uncharacteristic).

1. I have opinions about almost everything.
2. I like having responsibility for handling situations that require a lot of thinking.
3. It is very important to me to hold strong opinions.
4. I often prefer to remain neutral about complex issues.

A.5.2 TIPI for Big 5

Here are a number of personality traits that may or may not apply to you. Please indicate the extent to which you agree or disagree that these characteristics apply to you. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other. (On a scale from 1 to 7, with 1 = strongly disagree and 7 = strongly agree).

1. Extroverted, enthusiastic
2. Critical, quarrelsome
3. Dependable, self-disciplined
4. Anxious, easily upset
5. Open to new experiences, complex
6. Reserved, quiet
7. Sympathetic, warm
8. Disorganized, careless
9. Calm, emotionally stable
10. Conventional, uncreative

A.6 Trust in Automation Index

Many organizations now use algorithms to make forecasts. Some high profile examples include the use of statistics in baseball to choose players (Moneyball) or Nate Silvers use of statistics to predict elections. To what extent do you agree or disagree with the following statements about algorithms? (On a scale from 1 to 7, with 1 = strongly agree and 7 = strongly disagree). Given the variance in valence, items were coded so that the highest end of the response range (7) indicates *high* trust in automation and the lowest end of the range (1) indicates *low* trust.

1. Using algorithms increases the chances of organizations achieving their goals.
2. Using algorithms increases the effectiveness of organizations in making good decisions.
3. Using algorithms improves the output quality for organizations.
4. Using algorithms makes it more likely for organizations to make errors.
5. Modern organizations rely too much on algorithms to make decisions about the future.
6. Using algorithms is an effective way to overcome human biases.
7. When I am uncertain about something, I will trust the information from an algorithm in place of my own judgement.
8. When I am uncertain about something, I will tend to trust my own intuition and judgement over the information from an algorithm.

A.7 Base Relationships: Empirical Motivation

As an empirical motivation for our full study, we offer a short discussion of our base findings of relative influence of the treatment conditions in the experiment. We show the impact of advice from an algorithm or a judge relative to the baseline category of average past MTurk respondents for our two behavioral measures of trust in Table 3: advice weight and distance to advice. The strong positive impacts from the first model (column 1) for each condition suggest respondents are reacting to the advice, with the magnitude of the effect in the algorithm condition nearly twice that of the judge condition. Second, the pronounced negative effects in the second model (column 2) demonstrate the impact of the algorithm and judge treatments on reducing the distance between respondents' predictions and the advice-giver relative to the baseline category. Similarly, the effects are nearly doubled in the algorithm condition.

These results demonstrate two things. First, respondents were significantly more likely to change their evaluations based on the advice of "experts," whether human or machine-derived than they were to trust the "wisdom of the crowd." And second the algorithm condition is where the strongest effects are observed, suggesting respondents trust algorithms to a greater degree than advice from humans. This is an important finding by itself, and one we explore in greater detail in a separate paper.

Table 3. Experimental Impacts on Dependent Variables of Interest

	Dependent variable:	
	Advice Weight	Distance to Advice
	(1)	(2)
Algorithm Condition	0.134^{***} (0.013)	-6.563^{***} (0.864)
Judge Condition	0.073^{***} (0.013)	-3.812^{***} (0.857)
Constant	0.156^{***} (0.009)	27.353^{***} (0.608)
Observations	3,274	3,274
R^2	0.031	0.017
Adjusted R^2	0.030	0.017
Residual Std. Error (df = 3271)	0.305	20.113
F Statistic (df = 2; 3271)	52.076^{***}	29.117^{***}

*p<0.1; **p<0.05; ***p<0.01

Journal of Behavioral Data Science, 2022, *2* (1), 124–127.
DOI: https://doi.org/10.35566/jbds/v2n1/p8

Book Review: An Introduction to Nonparametric Statistics

Kévin Allan Sales Rodrigues*[0000−0003−4925−5883]

University of São Paulo, São Paulo, Brazil
kevin@usp.br

Book review of **An Introduction to Nonparametric Statistics** by John Edward Kolassa (2020). Chapman & Hall/CRC Press, Boca Raton. 224 pages. Price: USD 105.00 (print), 94.50 (e-book).

The book entitled *An Introduction to Nonparametric Statistics* (Kolassa, 2021) presents an exceptional introduction to nonparametric statistical techniques. To make it easier to read, I divided this book review into two sections: book description and comparisons with other books on nonparametric statistics.

1 Book description

According to the author, the book's target audience is graduate students in the applied statistical field. I find the book also suitable for graduate students researching applied mathematics and theoretical statistics. To make the best use of the book, the reader is expected to have a good knowledge of differential and multivariate integral calculus, matrix algebra, probability (mainly mean and variance of random variables), and classical inferential statistics.

A striking feature of the book is the logical and chronological chain of ideas. I classify the book as logical because it bothers to review the parametric statistics part corresponding to the nonparametric counterpart that will be covered immediately afterward. This logical approach favors learning because many undergraduate students study the parametric statistics before the nonparametric counterparts, and students can promptly relate what they previously learned in parametric statistics to what they study in nonparametric statistics. Thus, the book teaches nonparametric statistics more naturally and logically using the connection with the parametric methods. The first part of each chapter, except for two chapters: Chapter 1 (this is a leveling chapter of the book) and Chapter 8 (this is a chapter more focused on empirical probability density function graphing techniques), provides a brief review of the parametric technique equivalences to the nonparametric techniques that will be presented below.

I say that the book is chronological because it presents contents of nonparametric statistics in the same order as the articles that proposed such nonparametric techniques, favoring the understanding of the contents as the degree of complexity of the techniques increases gradually. For example, the author is constantly concerned with presenting the popular and traditional nonparametric tests: sign test, Mann-Whitney-Wilcoxon test, and Kruskal-Wallis test in chronological order, which are also included in other nonparametric statistics books.

The book has a brief introduction, 10 chapters and 2 appendices. The introduction leaves out something to be desired as to the fact that it does not list all the packages that will be used later in the text. It is good practice to list all packages used in the book, as it is helpful for the reader to prepare the computer environment in advance. It is also convenient to allow a university's computer lab with all the necessary packages already installed on the computers. For accessibility, I list here all the packages needed to run the R (R Core Team, 2022) code in the book: *MultNonParam*, *NonparametricHeuristic*, *devtools*, *BSDA*, *exactRankTests*, *DescTools*, *CvM2SL2Test*, *clinfun*, *muStat*, *crank*, *deming*, *Hotelling*, *ICSNP*, *KernSmooth*, *quantreg*, *MASS*, *boot*, *bootstrap* and *VGAM* [1].

Chapter 1 reviews the probability density functions and properties of random variables used throughout the book as well as the definitions and results of classical statistical inference (confidence intervals and hypothesis testing). Thus, this chapter is essential for the textbook to be self-sufficient, as far as possible, within the area of nonparametric statistics.

The content of nonparametric statistics starts from Chapter 2, which covers nonparametric inferential techniques for a single sample, such as the sign test.

Chapter 3 presents nonparametric inferential techniques for the case of two samples, and the theory behind the other tests and rank tests, is presented. Also, in Chapter 3, the widely used nonparametric tests such as Mann-Whitney and Mann-Whitney-Wilcoxon tests are presented; Mood's median test is presented for didactic purposes.

Chapter 4 extends the results presented to the case where we have three or more groups, explaining in detail the Kruskal-Wallis test. Nonparametric techniques analogous to the ANOVA (Analysis of Variance) of parametric inference are presented in Chapter 5.

Chapter 6 comments on the limitations of interpreting the Pearson correlation coefficient and aims to present alternative correlation measures to the Pearson correlation coefficient, which are the correlation coefficients proposed by Spearman and Kendall.

Chapter 7 presents techniques to perform multivariate nonparametric inference (context where observations contain more than two variables), such as population position parameter vector estimation and hypothesis testing about this same parameter vector.

[1] Note: some packages are already installed along with R or need to be installed from GitHub or CRAN.

Chapter 8 is the shortest in the book and presents techniques for estimating a sample's probability density function and plotting empirical probability density functions. The histogram graph is also presented in detail in this chapter.

Alternative regression techniques such as Kernel regression, Local regression, Isotonic regression, Quantile regression, and "Resistant regression" are presented in Chapter 9. When commenting on "Resistant regression", the author missed the opportunity to present basic results about the classes of M-estimators and even to comment on the existence of other classes of estimators (L-estimators, S-estimators). Knowing the specific terminology would be useful if the reader were interested to learn more about robust estimators. The topic of Splines is also covered in this chapter.

The final chapter introduces two resampling techniques: bootstrap and Jackknife. Several types of bootstrap are presented, as well as the advantages and disadvantages of each type of bootstrap.

Appendix A presents the codes in *SAS* equivalent to the codes presented in *R* throughout the course. Appendix B contains the codes in *R* used to generate the various tables and graphs presented throughout the text to support the author's argument. Thus, Appendix B makes the material presented in the book reproducible by any reader with basic knowledge of *R*.

Two packages of *R* are used extensively throughout the book, to illustrate the results of different tests and nonparametric techniques or to do data analysis, and they are *MultNonParam* e *NonparametricHeuristic*. The *MultNonParam* package is available on CRAN, and the *NonparametricHeuristic* package is available on GitHub. The book's author created both packages. As the package names suggest, *NonparametricHeuristic* is useful for teaching nonparametric statistics while *MultNonParam* is actually used in data analysis.

2 Comparisons to *Nonparametric Statistical Methods Using R*

An important component of the book is illustrating applications of the techniques to real data through free software *R*. This greatly enriches the text, as the text presents both theory and applications as well as provides the code in *R* and *SAS* so that the reader can carry out their own statistical analyses. Other books also follow this same approach of presenting code for nonparametric data analysis using *R*, such as *Nonparametric Statistical Methods Using R* (Kloke & McKean, 2015). I now compre the book *Introduction to Nonparametric Statistics* (INS) to the book *Nonparametric Statistical Methods Using R* (NSMUR). Below, I list the point-by-point similarities and differences between them.

2.1 Similarities

1. Both books illustrate the theory through applications using *R*.
2. Both books provide the code for the examples.

3. Both books are concise, INS has about 200 pages of content while NSMUR has about 250 pages of content.
4. Both books cover resampling techniques such as bootstrapping.
5. Both books cover the most widespread nonparametric techniques while having additional topics of its own. For instance, INS provides a brief introduction to the concepts of local regression, isotonic regression, and quantile regression. NSMUR has a chapter dedicated to techniques for survival analysis such as Kaplan-Meier estimator and log rank test as well as a chapter dedicated to cluster correlated data analysis.

2.2 Differences

1. Only NSMUR includes a chapter reviewing basic *R* topics.
2. Only INS presents the codes in *SAS*.
3. Only INS makes a chapter reviewing basic topics of probability and classical statistical inference.
4. NSMUR has more exercises than INS, which is an advantage for the reader who likes to fix learning through exercise solving.
5. Only INS features Jackknife.

The INS book integrates the examples and R codes with the text, making the reading very fluid. Compared to more traditional books like Hettmansperger (1984) and Hettmansperger and McKean (2011), it is notable that INS is more focused on the applied part of nonparametric statistics and offers more comments on *R* code. In summary, the INS and NSMUR books fulfill the role of introductory books on nonparametric statistics very well.

It would be interesting to see in a possible next edition of the INS an extension of the chapter dealing with alternative regression methods, approaching isotonic regression and quantile regression with a little more theoretical depth.

Acknowledgements

The author gratefully acknowledges CNPq (Brazilian National Council for Scientific and Technological Development) for the financial support by grant #141836/2020-2.

References

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. Krieger.

Hettmansperger, T. P., & McKean, J. W. (2011). *Robust nonparametric statistical methods*. CRC Press.

Kloke, J., & McKean, J. W. (2015). *Nonparametric statistical methods using r*. CRC Press Boca Raton.

Kolassa, J. E. (2021). *An introduction to nonparametric statistics*. CRC Press.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org`