**ORIGINAL PAPER**

# Advancing Organizational Science Through Synthetic Data: A Path to Enhanced Data Sharing and Collaboration

Pengda Wang[1] · Andrew C. Loignon[2] · Sirish Shrestha[2] · George C. Banks[3] · Frederick L. Oswald[1]

## Abstract

The importance of data sharing in organizational science is well-acknowledged, yet the field faces hurdles that prevent this, including concerns around privacy, proprietary information, and data integrity. We propose that synthetic data generated using machine learning (ML) could offer one promising solution to surmount at least some of these hurdles. Although this technology has been widely researched in the field of computer science, most organizational scientists are not familiar with it. To address the lack of available information for organizational scientists, we propose a systematic framework for the generation and evaluation of synthetic data. This framework is designed to guide researchers and practitioners through the intricacies of applying ML technologies to create robust, privacy-preserving synthetic data. Additionally, we present two empirical demonstrations using the ML method of generative adversarial networks (GANs) to illustrate the practical application and potential of synthetic data in organizational science. Through this exploration, we aim to furnish the community with a foundational understanding of synthetic data generation and encourage further investigation and adoption of these methodologies. By doing so, we hope to foster scientific advancement by enhancing data-sharing initiatives within the field.

**Keywords** Synthetic data · Machine learning · Open science · Data sharing

Today's era of open science critically involves the sharing of data to strengthen scientific understanding, evidence-based practice, and policy making (Banks et al., 2019; Nosek et al., 2015). Yet, organizational researchers and other social scientists are often hesitant to share their data (e.g., Hardwicke et al., 2021; Towse et al., 2020; Vanpaemel et al., 2015; Wicherts et al., 2006). Vanpaemel et al. (2015) offered a telling example, by reporting that out of 394 data requests made across four American Psychological Association (APA) journals, a mere 150 researchers (38%) ultimately shared their data. The open science movement hopes to reverse this trend, where journals and the federal government are beginning to

encourage, strongly recommend, or even require the sharing of data and materials, as found in the Transparency and Openness Promotion (TOP) guidelines of the Open Science Framework (Nosek et al., 2015). In the scholarly context, data-sharing activities seek to boost transparency, by allowing for analytic reanalysis and reproducibility, bolstering the credibility of scientific research (Pew Research Center, 2019; Towse et al., 2020). Moreover, the availability of open data allows for additional extended analyses, in addition to much greater flexibility and power when conducting meta-analyses across studies that provide the raw data.

Compounding this problem are many challenges to sharing data in organizations. Ensuring anonymity is one of them, given that organizations entrusted with employee and job applicant data seek to keep the identity of individuals not only anonymous but completely private. Other reasons are that organizations stand to gain a competitive advantage from the data they glean insights from and do not share. Additionally, they avoid the risk of reputational damage by not sharing data that might reveal unfavorable information about the organization. Keeping these types of ethical, privacy, and proprietary issues well in mind, we believe that *synthetic data* could offer one promising solution to

---

surmounting at least some of these hurdles, for those organizations seeking to be thought leaders by participating in scientific activities in the aforementioned spirit of data sharing and open science.

Jordon et al. (2022) defined *synthetic data* as "data that have been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)." If one were to rely on this definition, Monte Carlo simulation data could also be considered synthetic data. For example, researchers can simulate data for subsequent *t*-tests, ANOVAs, regression, SEM, and other forms of modeling and hypothesis testing (e.g., generating simulation data using the mvnorm function within the MASS package in R; Venables & Ripley, 2002). But simulations are based on specifying underlying parameters and distributions beforehand (e.g., a multivariate normal variance–covariance matrix and its associated mean structure). Thus, simulations generate data based on known parameters and attendant distributional assumptions; these data are then analyzed to see whether those parameters and assumptions can be well approximated by sample-based estimates.

In this paper, our definition of synthetic data slightly differs from that of Jordon et al. (2022). We emphasize the distinction between synthetic data and simulation data: while simulation data is generated based on multivariate distributional assumptions, synthetic data are generated by learning from the multivariate distribution of the original dataset itself. The goal is for analyses performed on synthetic data to closely mirror those performed on the original data. A synthetic dataset therefore serves as an alternative to the original dataset, allowing one to mirror the original data in its realism and complexity while still ensuring data privacy and protection for the original data owners (Raghunathan, 2021). With these advantages, synthetic data can promote open data sharing in organizational research, encouraging academic-practitioner research partnerships that align with the recent and strengthening development of open science initiatives in the field (Castille et al., 2022).

At the same time, despite the strengths of synthetic data, it is essential to highlight that their security and substitutability are critically influenced by many important design factors. For example, for synthetic data to serve as a reasonable substitute for the original data, researchers need to determine their intended use, evaluate the efficacy and effectiveness of synthetic data generation techniques closely, and ensure stringent data anonymity and otherwise uphold ethical, privacy, and proprietary standards.

The primary goal of our paper is to develop a systematic framework that guides organizational researchers through the synthetic generation and evaluation processes, as we are unaware of any such resources available to organizational scientists. This framework helps readers consider, generate, interpret, and use synthetic data from a broad perspective.

Through two specific examples, we will demonstrate how to apply machine learning (ML) generative models to learn from a given dataset. Although the ML method we explain and use is called generative adversarial networks (GANs), readers should know there are many other ML options for generating synthetic data. However, covering those here would detract from the central goal of our paper. Based on this framework, we also provide two empirical demonstrations. Ultimately, our goal is not to serve as a definitive resource, but rather to provide organizational scientists and practitioners with a general understanding and rationale behind generating synthetic data, hoping they will be motivated to pursue synthetic data methodologies further and expand upon our efforts.

We also discuss ideas for navigating potential ethical dilemmas and risks associated with developing and applying synthetic data (Porter, 2008; United Kingdom Statistics Authority, 2022). Taken as a whole, our research invites the broader integration of synthetic data within the organizational sciences, helping to unlock organizational data that has remained restricted, inaccessible, and thus ultimately lost over time (Wicherts et al., 2006). Ultimately, as more high-quality synthetic data are shared that strikes an appropriate balance between accessibility and privacy, we envision both organizational science and evidence-based practice to benefit greatly.

## Synthetic Data: Answering Key Questions

The term "synthetic data" is a relatively unfamiliar concept for most organizational scientists. When encountering a new research method, four main questions usually need answers: *What is this method? Why should we use this method? How does this method differ from previous methods? How should we implement this method?* Below, we will address each of these four questions in turn as they pertain to synthetic data.

### What Are Synthetic Data?

In this paper, we define *synthetic data* as artificially generated data designed to emulate the original data as closely as possible without revealing actual observations in that data. The purpose is to provide an alternative to original data in situations where using original data is impractical, poses privacy concerns, is disallowed (e.g., organizationally, legally, and ethically), or is otherwise restricted (Fonseca & Bacao, 2023; Jordon et al., 2022).

Consider an original dataset containing 1000 data points. The corresponding synthetic dataset will also consist of 1000 data points. However, it will result from a process that begins with random noise and iterates to approximate the original dataset (e.g., GANs). This iterative process ensures that the two datasets do not exactly overlap, thereby

protecting the privacy of the data. At the same time, both datasets will have similar multivariate distributional properties, ensuring that the synthetic dataset can substitute for the original dataset with very similar analytic results, no matter what analysis is used.

For example, consider a dataset containing detailed performance metrics for all NBA players in a season, including well-known players like LeBron James. The original dataset might include precise statistics such as points per game, assists, rebounds, and other performance indicators for each player (e.g., value over replacement). If we generate a synthetic dataset based on these original data, the new dataset will maintain the overall distribution of these performance metrics—such as average points per game or the range of assists across players—without directly replicating the statistics of any specific individual, including LeBron James. When generating the synthetic data, data obfuscation will be achieved by starting with noise and learning each player's metric features from the original dataset. The aim is that while individual data points are perturbed and become difficult to distinguish, the overall multivariate distribution characteristics are preserved, thereby maintaining the statistical properties of the original data while hiding individual data points. As a result, the synthetic data can be shared freely and used for analyses requiring statistical properties similar to the original data, with the added benefit of protecting the privacy of actual players.

It is important to note that original datasets typically comprise quantitative observations sampled from a larger population. Consequently, synthetic data is intended to imitate the multivariate distributional characteristics of the original data. As a result, the quality of synthetic data is closely tied to the quality of the raw data. In other words, if the original data contains non-representative samples, measurement error, extensive missing data, and other undesirable qualities (e.g., small sample sizes), then the synthetic data will necessarily reflect that, as the saying goes: "garbage in, garbage out."

## Why Would One Synthesize Data?

Privacy issues have always been an obstacle to data sharing with researchers (Gabriel & Wessel, 2013; Leavitt, 2013). Rubin (1993) highlighted the privacy risks of sharing microdata (i.e., information about individual transactions) and the complex legal issues it presents. He suggested using imputed synthetic data as a solution, where confidential information in a dataset is replaced with estimates generated from an imputation model, using the same background variables without disclosing the original sensitive data. Whereas imputations require assuming and estimating a model that underlies the data, ML-based methods can learn the model from the data themselves, thus improving upon Rubin's ideas. We will illustrate the use of one type of ML generative model, generative adversarial networks, or GANs (Goodfellow et al., 2014, 2020). The GAN procedure begins as noisy data but then shapes itself into a synthetic dataset as it learns its distributional form from an actual dataset. Therefore, a synthetic dataset based on a GAN does not contain any actual individual data.

Synthetic data primarily serves as a substitute for original data, particularly when the original data cannot be easily shared due to privacy concerns or other restrictions, such as the low rates of individual-level participant data sharing in the medical and health sciences (Hamilton et al., 2023). The issue of data availability has been persistent. For example, Alsheikh-Ali et al. (2011) found that high-impact journals do not always ensure the public availability of research data, with only 9% depositing full primary raw data online. Wicherts et al. (2006) found that 73% of scholars did not share data. Hardwicke et al. (2021) reported that even with a mandatory data-sharing policy in the journal *Cognition*, around 20% of researchers still withheld their data. Tedersoo et al. (2021) observed that although there has been some progress in recent years, data availability and the willingness to share data remain very low. Moreover, based on our own experiences and discussions with colleagues, we have consistently encountered significant challenges in securing organizational data that can be freely shared with others. While we are often permitted to publish findings derived from organizational data—typically in collaboration with organizational partners—the raw data remain inaccessible. This restriction applies not only to external researchers who might wish to verify or build upon our findings but also to peer reviewers assessing the rigor of our work. The inability to share data creates a substantial barrier to transparency and reproducibility in organizational research.

We believe that synthetic data can serve as a method to promote data sharing, and some studies have already adopted this approach. For example, Nowok et al. (2017) demonstrated that this method had been used in some UK longitudinal studies for data-sharing purposes. By using synthetic data as a substitute for original data, researchers can partially address issues like those mentioned by Couture et al. (2018), where data that cannot be shared initially may become nearly irretrievable later on, making it almost impossible to recover if it is not published (e.g., if the first author cannot be found, or if they retire or change affiliations). Colavizza et al. (2020) also indicated that articles containing data tend to have a greater impact, and providing synthetic data can help promote the dissemination of the article to some extent. We hope this paper can serve as a starting point for the use of synthetic data in organizational research, provide a general understanding and rationale behind generating synthetic data, and call for more thought and research on synthetic data within the field.

## How Do Synthetic Data Compare to Data Generated by Existing Methods?

In order to highlight the features of synthetic data, we compare it with similar but distinct data from two existing methods: anonymizing data and simulating data.

### Anonymizing or De-identifying Data

Anonymization is the process of removing or altering personal identifiers from original data so that the individuals described by the original data become anonymous. Often, it involves changing or removing those variables that could directly or indirectly lead to identifying individuals in the dataset. Anonymized data have some use, but their value can be compromised due to having to remove, combine, or alter key features of the original dataset. For example, removing demographic information directly results in the inability to study these variables; binning the data as a solution to keeping individual demographics anonymous would limit the variability of the data (Cohen, 1983) and limit the appropriate analyses and inferences that can be made as well. By contrast, synthetic data offer several advantages over simply anonymizing the original data. When produced well, synthetic data do not contain the original data of any individuals, thus offering a higher level of privacy than simply removing personal information from the original dataset. Likewise, synthetic data do not require removing or combining variables, thus retaining the essential qualities of the original data. In contrast with the previous example, demographic information can often still be retained as part of a synthetic dataset.

When the anonymization of data does not impact its subsequent analysis, anonymized data are generally more useful than synthetic data. For example, if demographic information in the original data is not central to the research, then best practice might be to remove this information before sharing the data. However, the key issue that, in fact, motivates this paper is that proprietary concerns, privacy regulations, and other legal and ethical imperatives may restrict the sharing of organizational data or certain important parts of the data, which affect subsequent analysis. It is in these situations where synthetic data serve as an important alternative, to allow for data sharing, replicating past analyses on the original data, and conducting new analyses to gain new insights from the data that would otherwise be inaccessible.

### Simulating Data

Often researchers are unsure of the distinction between synthetic data and simulated data. Monte Carlo simulations, for example, are generated based on parameter estimates (e.g., a range of correlations or specific values, say estimates from a meta-analysis). These estimates are then incorporated into prespecified models (e.g., regression, ANOVA) to generate data with modeled distributional assumptions (e.g., multivariate normality with prespecified variances and covariances, specific proportions in each group, psychometric reliability and validity estimates, subgroup mean differences). Although this simulation-based approach affords the user tremendous flexibility and control over the data-generating mechanism, the simulated data (and its underlying model and parameters) may or may not be generalizable to the real world. By contrast, synthetic data are generated using an actual dataset to develop the best approximation of that dataset using ML models (in our case, GANs). Because synthetic data are derived from the original data, there are no distributional assumptions, and results are intended to reflect and thus generalize to the original dataset.

In conclusion, although simulated data offers versatility and control for theoretical exploration, synthetic data provide a reflection of real-world data (with the good qualities and the bad), making it a powerful tool for substituting original data when data sharing is restricted due to privacy issues. Researchers should carefully consider the context and goals of their study when choosing between these two types of data, balancing the need for theoretical exploration with the desire for practical applicability.

## How Can a Researcher Synthesize Data?

To facilitate the generation, use, and adoption of synthetic datasets within organizational research, we outline a general process to guide organizational researchers (see Fig. 1). In general, this process consists of two major steps, data generation, and evaluation, which we have color-coded accordingly. Next, we review each of these steps along with the specific considerations within each.

### Examine Original Data and Determine Intended Use of Synthetic Data

The first step of data generation is to examine the original data and determine its intended use for synthesis. The features of the original data, and the intended use of the synthetic data, are both key factors in understanding and guiding the subsequent steps of the process.

As previously mentioned, synthetic data are generated by learning from the multivariate distribution of the original dataset itself. However, different ML models have their own strengths and adaptability in learning and capturing various features. For example, the Gaussian Copula models work particularly well for data that are approximately normal or can be made approximately normal through transformations. On the other hand, GANs work better for capturing complex,

**Fig. 1** Flow chart summarizing the general process steps for synthesizing data. *Note*: GANs, generative adversarial networks; VAEs, variational autoencoders
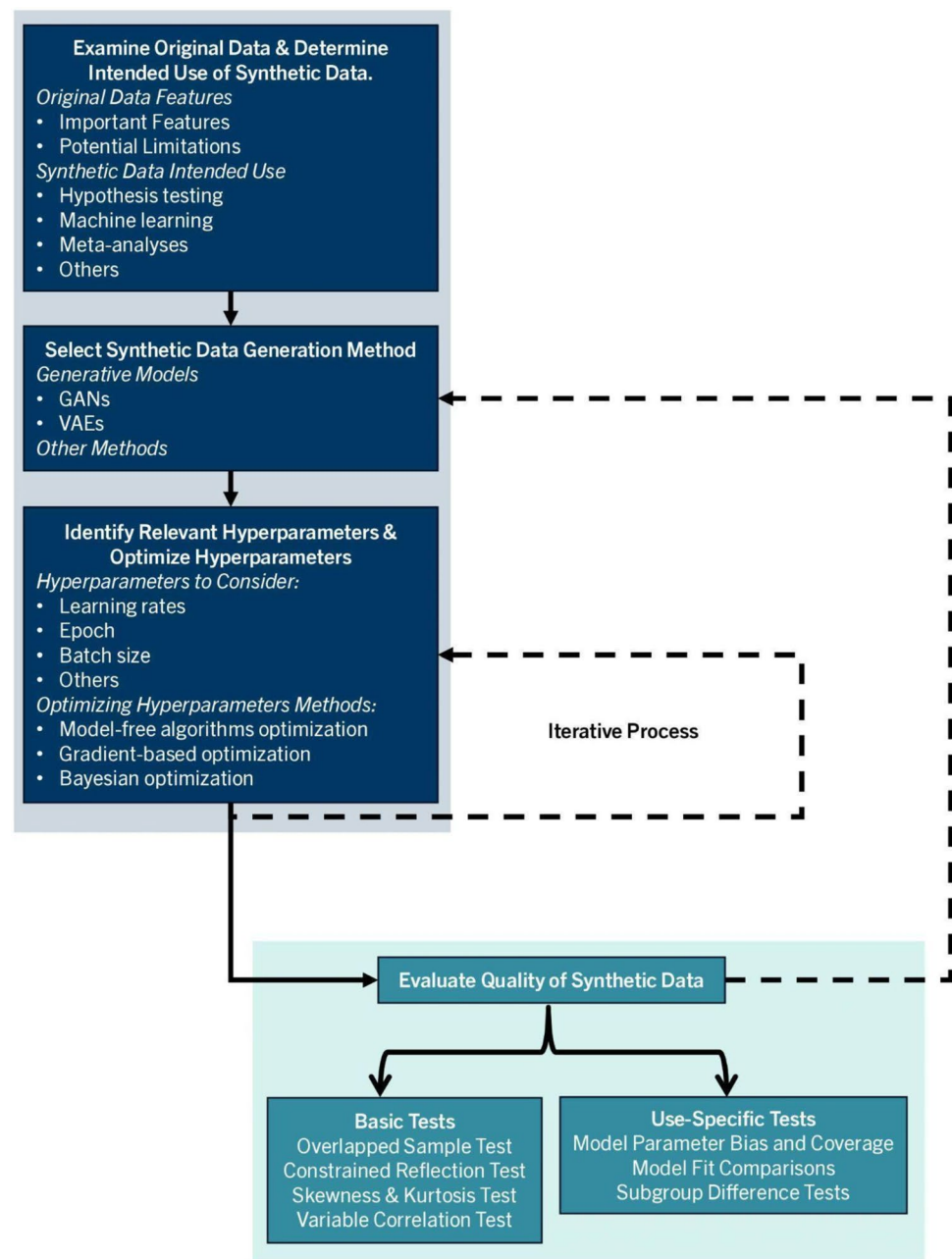


high-dimensional distributions, making them suitable for more complex data.

For different types of original data, the characteristics of different models determine how well they perform in learning them and generating corresponding synthetic data. Likewise, the complexity of relationships may limit the effectiveness of synthetic data methods. Greater complexity in the relationships between variables makes generating accurate synthetic data more challenging, leading to increased computational complexity. Besides that, the intended use of the synthetic data also influences our choice of generative model selection and evaluation. Different intended uses mean that

the synthetic data need to focus on learning different attributes of the original data, and different generation methods may be better suited for producing specific attributes.

We foresee at least three possible uses of synthetic data. First, researchers who cannot share their original data may generate and share synthetic data so that others can approximate and perhaps extend their original analyses (e.g., hypothesis tests, factor analyses, structural equation models). Note that the original data may not have satisfied the assumptions of the statistical models used (e.g., homoscedasticity for regression analysis), and results may not have been statistically significant. Nonetheless, the goal for

synthetic data is to be as close to the original data as possible, so that the properties of synthetic data and any analytic results conducted on them are as close to those from the original data as possible.

Second, researchers may generate synthetic data to help build, test, and train ML models. Synthetic data present an opportunity to access a version of the original data that previously could not be used for training, due to privacy concerns. Thus, many existing applications generate synthetic data as input into ML models whose primary goal is often similar levels of prediction or classification as with the original data, versus obtaining similar parameter estimates of a statistical model. It should be noted that researchers have generated artificial data to supplement an original dataset, increase its diversity, and train fairer ML models as a result (e.g., Feldman et al., 2014; Zhang et al., 2016). To be clear, this type of "de-biased synthetic data" is different from the synthetic data discussed in this paper, which are designed to emulate the original dataset as closely as possible without any extensions of this nature.

Third, synthetic data may be generated to inform meta-analyses or integrative data analyses that summarize relationships or effect sizes within the existing literature (Curran & Hussong, 2009). In this application, synthetic data would allow researchers to more readily share their original data, which may help inform more nuanced summaries of a given field (e.g., just as item-level meta-analyses have, such as Carpenter et al., 2016).

For example, if the intended use of synthetic data is to help build, test, and train ML models, we will be more concerned with how closely the distribution of the synthetic data matches the original data. This includes the marginal and joint distributions of features, the correlation structure among features in the original data, and the proportions of various sample types in the original data. Given that GANs perform well in capturing complex distributions and correlations in data, using GANs might be a good choice.

### Select Synthetic Data Generation Method

Once the researcher has a sufficient level of familiarity with the original data and has a clear goal for the use of synthetic data, they can select a generation method to use. We will mainly introduce methods based on GANs (Goodfellow et al., 2014, 2020). GANs consist of two neural networks—the Generator and the Discriminator—that work together to produce high-quality synthetic data (see Fig. 2). The Generator creates data that are rewarded for generating and mimicking an original dataset, while the Discriminator is rewarded whenever the data are correctly judged to be inauthentic. The Generator and Discriminator will also include additional settings to prevent generating samples that are identical or overly similar to the original

data and to reject any duplicates (e.g., data deduplication; Xia et al., 2016). Through this adversarial process, both the Generator and Discriminator iteratively improve, enabling them to work together to produce data that serves as a substitute for the original, while preserving privacy.

GANs can be combined with various statistical principles and other models to form more complex and powerful methods, such as Conditional Tabular GAN (CTGAN; Xu & Veeramachaneni, 2018; Xu et al., 2019) and Copula GAN (Kamthe et al., 2021). Readers should keep in mind that each method has its strengths and limitations, and what works well for one dataset or application might not be suitable for another. The right tool for the right job understands and selects the appropriate synthetic data generation method to ensure the preservation of original data characteristics and meet the intended use.

### Identify Relevant Hyperparameters and Optimize Hyperparameters

Most ML-based synthetic data generative models, including GAN-based models, allow researchers to fine-tune various hyperparameters to help ensure that the synthesized data exhibits preferred qualities. For example, there are several options available, ranging from epoch (i.e., how many times the training data are used to train the model), batch size (i.e., the amount of data fed into the model during each training iteration), and learning rates (i.e., the speed at which network parameters are updated during training). Appendix provides more details on the differences between parameters and hyperparameters, as well as several hyperparameters and common methods for optimizing hyperparameters.

Interestingly, the tuning or adjusting of these hyperparameters is highly related to computational resources and time, which is also a consideration for generating synthetic data. Generally speaking, allowing the model to learn from and emulate the original data, spending more training rounds, and using more complex models results in better outcomes (see Appendix). However, this also means investing more time and money. Therefore, setting a minimum target performance standard is particularly important. The minimum target performance standard refers to the lowest acceptable threshold during evaluation, which means it essentially meets or nearly meets the intention of generating synthetic data.

### Evaluate the Quality of Synthetic Data

An important question about synthetic data is how to evaluate its quality relative to the original data. Two important evaluation criteria are *security* and *substitutability*. We categorized the evaluation tests into "basic tests" and "use-specific tests."
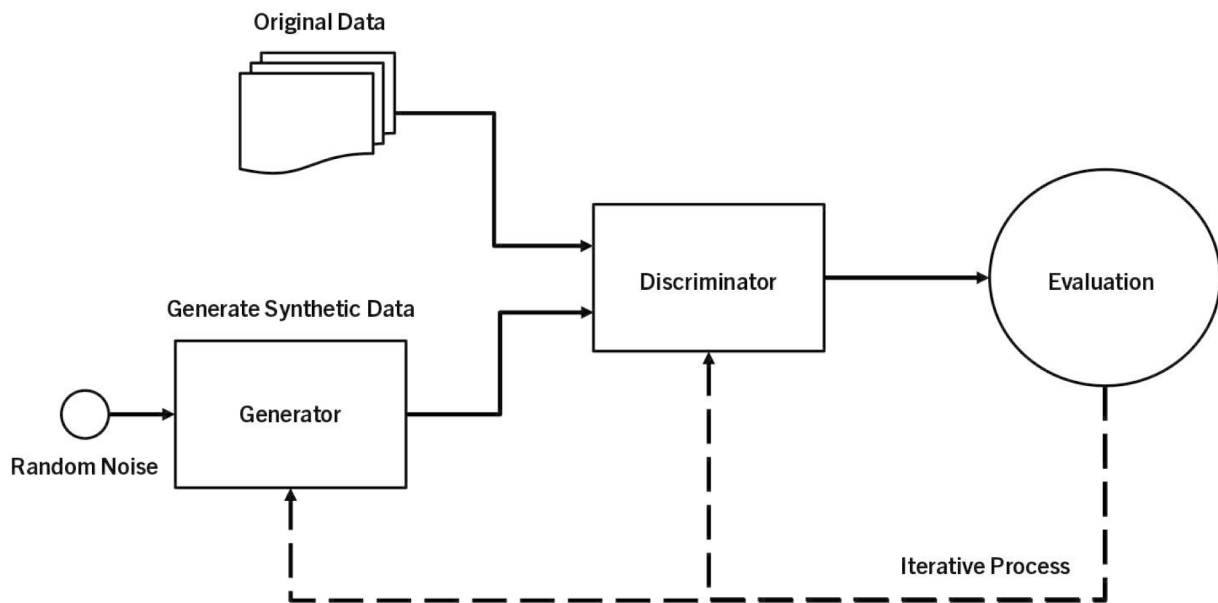
*GANs Workflow*

**Fig. 2** GAN workflow. *Note*: All GAN-based models share a typical workflow consisting of two main components: the Generator and the Discriminator (or Critic). The Generator's primary role is to produce authentic data so the Discriminator cannot differentiate it from the original data. Conversely, the Discriminator's job is to identify whether the input data are actual or fabricated by the Generator. Both components enhance their capabilities through their competition, ultimately leading to data that more closely resemble the original dataset

Basic tests provide the fundamental checks for synthetic data, including descriptive statistics such as means and correlations, which are important for any synthetic data. Use-specific tests, on the other hand, focus on the performance of synthetic data in target applications. For instance, if the synthetic data are intended for training ML models, the results should be comparable to those obtained with the original data.

**Synthetic Data Quality: Basic Tests** Four basic tests provide a general assessment of the security and substitutability of synthetic data.

*Overlapped Sample Test:* This test checks the uniqueness of the synthetic dataset, an important step to prevent data replication. Specifically, it assesses the proportion of cases in the original and synthetic datasets that share the same value. Lower values indicate less overlap, thereby reducing the risk of security concerns.

$$Percentage = \frac{Number of same values}{Total number of values} \times 100$$

*Constrained Reflection Test:* This test checks how well each generated value conforms to the range of the original variables. By calculating the minimum, maximum, mean, and standard deviation in both the original and synthetic datasets, as well as effect sizes capturing the differences in these measures, one can demonstrate whether there is sufficient alignment and consistency.

*Distribution Kurtosis and Skewness Test:* This test checks the distribution characteristics of the datasets, allowing one to understand the asymmetry and the "tailedness" of the data distribution, which can be important in assessing the quality of the synthetic data. By analyzing the kurtosis and skewness, one can determine how closely the synthetic data mimic the original dataset in terms of distribution shape.

*Variable Correlation Test:* This test checks the correlation between datasets' variables, which is important for gauging the substitutability of the synthetic data. The datasets reveal interrelationships among the variables. We scrutinized how these correlations were shown in each generated dataset and computed the mean of these values.

**Synthetic Data Quality: Use-Specific Tests** Along with the basic tests of synthetic data quality, researchers must also provide evidence that synthetic data provide reasonable approximations of the parameters that will be of most interest or relevance to the eventual end-user. Some examples for each of the potential uses include as follows:

*Hyphothesis Testing*: Researchers intending to generate and share synthetic data as a means of allowing others to approximate and extend their work may be particularly concerned with whether the synthetic

data reproduce the important parameters within their models. Within organizational research, these parameters often include factor loadings, regression coefficients, and covariances. Thus, demonstrating that these parameter estimates fail to exhibit bias (i.e., minimal differences between the values obtained in the original data and the synthetic data), have adequate coverage (i.e., overlapping confidence intervals for the synthetic and original data), and yield similar conclusions for significance testing across the synthetic and original data can help assure future users of the synthetic data of the reproducibility of the results.

*Machine Learning:* When using synthetic data to build, train, and test ML models, the most important criterion may be the model fit. With ML models, researchers are often less concerned with the value of specific parameter estimates and are seeking to identify efficient and accurate means by which they can classify cases or predict future values. Examples of model fit estimates, then, that a researcher could use to show that the original and synthetic data are returning comparable results include $R^2$, mean square error (MSE), and mean absolute error (MAE).

*Meta-analyses:* If using synthetic data to inform meta-analyses or integrative reviews, what likely becomes paramount is whether the effect size(s) obtained from synthetic data are similar to those from the original data, for all practical purposes, contributing to a meta-analysis in the same way as the original study would. Moreover, synthetic data provided across studies might be combined to allow for more refined analyses of moderator effects through multilevel modeling, going beyond a moderator analysis of effect sizes in a traditional meta-analysis.

These examples above—hypothesis testing, ML, and meta-analysis—illustrate that evaluating the quality of synthetic data involves reasoned demonstrations that their synthesized data are well-suited for their intended purpose(s). One can therefore never "collect stamps" (Landy, 1986) and say categorically that synthetic data are always as good as the original dataset. Potential limitations of synthetic data, along with recommendations for data sharing, are further discussed in the discussion section.

### Feedback Loop from Data Evaluation Back to Data Generation

It is likely that during this process researchers will obtain poor results with one, or perhaps several, of the basic or use-specific criteria. As such, researchers may adopt a pareto-optimization approach when evaluating their synthetic data where they seek

to reach defensible thresholds across multiple criteria and prioritize those that are most important given the intended use of the data. As depicted in Fig. 1, to achieve such optimization, we anticipate that researchers will have to cycle back to some of the initial data generation steps after evaluating the synthetic data that are initially obtained.

This represents a broader process beyond just tuning hyperparameters. The iterative loop for hyperparameter optimization produces good results for a model within a specific range, guided by a statistical loss function (see Appendix). The loss function is a fundamental measure for assessing synthetic data. It indicates the degree of inconsistency between the model-generated values and the original values (e.g., MSE, MAE). The loop from data evaluation back to data generation is an overall cycle of generating synthetic data. Through a more detailed evaluation, we may decide to adopt different models to improve the quality of the generated synthetic data.

### Additional Considerations

Before proceeding, there are a few aspects of this workflow that are worth noting. First, we acknowledge that the techniques available for synthesizing data are developing rapidly (e.g., data generation models and hyperparameter optimization). Thus, the specific approaches used within some of these steps will likely need to evolve as improved techniques emerge. Second, we do not presume that this process is solely the responsibility of the researchers (i.e., individual synthesizing the data), but should also be something that end-users consider before incorporating synthesized data into their own work. That is, the end-user of the synthesized data should use the proposed workflow to consider whether there is sufficient information available to determine the rigor and quality of the synthetic data given its stated purpose. Third, consistent with broader recommendations regarding transparency and open science, this workflow will be most effective if researchers disclose the steps used in the synthetic data generation process and provide a comprehensive summary of the results of the tests of the evaluation criteria.

## Empirical Demonstration of Synthetic Data and Two Research Questions

As mentioned before, a primary benefit of a synthetic dataset is that it contains analytically interpretable raw data for conducting a wide range of analyses without sacrificing the privacy of the individuals contributing to the dataset. Thus, an important question is assessing the extent to which a given synthetic dataset can sufficiently substitute for the original dataset that generated it. Addressing this question is important to support the broader use of synthetic datasets in organizational research. With the goal of trying to increase the synthetic data's similarity to the original dataset, this must

be accomplished in a manner that safeguards the sensitive and personal information of respondents. Optimal synthetic data should precisely retain the statistical attributes of the original data while minimizing any resemblance, patterns, or details that could inadvertently jeopardize the anonymity and security of participants and their data, as found in the original dataset.

Motivated by these goals, we present two empirical demonstrations of synthetic data that address the following research questions:

*Research Question 1 (RQ1):* To what extent do synthetic data possess sufficient <u>substitutability</u>, as captured original data patterns, to serve as a viable alternative to raw datasets?

*Research Question 2 (RQ2):* To what extent do synthetic data possess sufficient <u>security</u> and limit the percentage of duplicated cases with the original dataset?

The analytic codes for both studies are available on GitHub: https://github.com/wpengda/SyntheticData_OrganizationalScience. We used the Python SDV package (Patki et al., 2016) to generate synthetic data. Data were analyzed using R, version 4.4.0 (R Core Team, 2024), as well as the psych package version 2.4.6.26 (Revelle, 2024).

## Study 1: Synthetic Data in the Context of Assessment

Our first study used the General Aptitude Test Battery (GATB) dataset (U.S. Department of Labor, 1970). The GATB is an expansive assessment tool that provides nine cognitive aptitude scale scores on the basis of 12 tests thought relevant to the prediction of job performance. These aptitudes are General Learning Ability (G), Verbal Aptitude (V), Numerical Aptitude (N), Spatial Aptitude (S), Form Perceptual (P), Clerical Perception (Q), Motor Coordination (K), Finger Dexterity (F), and Manual Dexterity (M) (see Table 1). In addition to GATB scores, the dataset contains job performance criterion scores and detailed job descriptions. Thus, the GATB dataset is akin to others commonly employed in the context of assessment, hiring, and selection (Bemis, 1968; Vevea et al., 1993).

### Participants and Procedure

The entire GATB dataset consists of 40,489 employees in various occupations who took the GATB. For the current study, just for the purpose of an example, we focus on a specific occupation: Gambling Dealers (O*NET SOC Code 39–3011.00), which comprises 1056 employees. The sex distribution was 57.1% male and 42.9% female. In terms of ethnicity, it is predominantly White (85.8%), followed by Black (11.8%), with the remaining 2.4% representing other ethnic groups. The measures of the aforementioned nine aptitudes identified in the GATB data can also be seen in Table 1.

### Analyses

To generate the synthetic data for the GATB dataset, we followed the process described earlier and depicted in Fig. 1. First, we examined the original data and found that some variables were skewed and leptokurtic (i.e., non-normal). Subsequently, we hypothesized that the synthetic data's intended use was to assist in building, testing, and training ML models. We focused on the standardized scores for each of the nine aptitudes identified in the GATB data and also included the standard supervisory rating scale as the performance criterion (CRFIN) in the analysis. Essentially, the final performance criterion was predicted by these nine ability indices. Additionally, the GATB dataset includes demographic variables such as sex and race. Therefore, we also wanted to explore whether the original data and synthetic data produced similar results under the same demographic variables.

Based on these purposes, we selected the generative model CTGAN as the method for generating synthetic data and adjusting multiple hyperparameters, such as the number of epochs. As shown in Fig. 1, optimizing hyperparameters is an interactive process; for the sake of parsimony, we only present the final results here (i.e., batch size = 128; epoch = 5000).

### Study 1 Results

#### Basic Test

Overlapped Sample Test.

We began by conducting an overlapped sample test to ensure that the synthetic data did not inadvertently create potential security or privacy risks. This test indicated that there was no overlap in data cases between the synthetic and original data. This suggests a certain level of security in the synthetic data.

Constrained Reflection Test.

Next, constrained reflection test findings are presented in Table 2. This table encompasses each variable's minimum and maximum values, median, mean, standard deviation, and Cohen's *d*. Cohen's *d*, which quantifies the effect size difference between original and synthetic data, was less than the small effect threshold of 0.20 in all cases. Some people may argue that even a tiny Cohen's *d*, like 0.10, reflects meaningful differences. However, we believe that if the goal is to detect any difference, regardless of size, then a small effect size like 0.10 may warrant attention. On the other hand, if the goal is to identify changes of

**Table 1** General Aptitude Test Battery (GATB): aptitudes and corresponding subtests

| Aptitude | Definition | Subtest(s) |
|---|---|---|
| **G**-General Learning Ability | The ability to "catch on" or understand instructions and underlying principles; the ability to reason and make judgments. Closely related to doing well in school | Part 3-Three-Dimensional Space<br>Part 4-Vocabulary<br>Part 6-Arithmetic Reasoning |
| **V**-Verbal Aptitude | The ability to understand the meaning of words and to use them effectively. The ability to comprehend language, to understand relationships between words, and to understand the meaning of whole sentences and paragraphs | Part 4-Vocabulary |
| **N**-Numerical Aptitude | Ability to perform arithmetic operations quickly and accurately | Part 2-Computation<br>Part 6-Arithmetic Reasoning |
| **S**-Spatial Aptitude | Ability to think visually of geometric forms and to comprehend the two-dimensional representation of three-dimensional objects. The ability to recognize the relationships resulting from the movement of objects in space | Part 3-Three-Dimensional Space |
| **P**-Form Perception | Ability to perceive pertinent detail in objects or in pictorial or graphic material. Ability to make visual comparisons and discriminations and see slight differences in shapes and shadings of figures and widths and lengths of lines | Part 5-Tool Matching<br>Part 7-Form Matching |
| **Q**-Clerical Perception | Ability to perceive pertinent detail in verbal or tabular material. Ability to observe differences in copy, to proofread words and numbers, and to avoid perceptual errors in arithmetic computation. A measure of the speed of perception which is required in many industrial jobs even when the job does not have verbal or numerical content | Part 1-Name Comparison |
| **K**-Motor Coordination | Ability to coordinate eyes and hands or fingers rapidly and accurately in making precise movements with speed. Ability to make a movement response accurately and swiftly | Part 8-Mark Making |
| **F**-Finger Dexterity | Ability to move the fingers, and manipulate small objects with the fingers, rapidly or accurately | Part 11-Assemble<br>Part 12-Disassemble |
| **M**-Manual Dexterity | Ability to move the hands easily and skillfully. Ability to work with the hands in placing and turning motions | Part 9-Place<br>Part 10-Turn |

*Note:* Adapted from the U.S. Department of Labor (1970). As cited in Kato and Scherbaum (2023)

a certain magnitude or greater—such as when we want to create synthetic data that is not the same as the original—then such a small effect size might be considered negligible.

Distribution Kurtosis and Skewness Test.

Table 2 also reports the skewness and kurtosis values. The kurtosis for both the original and synthetic data sets showed a high degree of similarity across variables, all close to three (Mesokurtic), indicating a nearly normal distribution. Although the skewness was also closely matched for most variables, exceptions such as V and P exhibited opposite signs, suggesting a divergence in their asymmetry. However, given that the skewness values were within the $\pm 0.5$ range, we could infer that the distribution of these variables was approximately symmetric. This near-symmetry, coupled with the absence of pronounced features, might explain why the synthetic data failed to capture the more nuanced trends present in the original data.

Variable Correlation Test.

We present the findings of the variable correlation test in Table 3. Most of the correlations between the original and synthetic data aligned closely, with only a few exhibiting minor differences. The mean absolute difference in the magnitude of correlation coefficients across the two datasets was 0.06 (*SD* = 0.05, min = 0.00, max = 0.22), indicating strong similarity at this level of correlation. Figure 3 provides a

histogram summarizing these differences, 76% of the correlation estimates exhibited differences less than or equal to 0.10. Significantly correlated variables in the original dataset generally retained their significance in the synthetic dataset. Instances of non-significant correlations or those that diverged in direction compared with the original data maintained absolute values below 0.22. This also offered a plausible explanation for certain discrepancies observed in the synthetic data, where weak correlations were infrequent.

## Use-Specific Tests

Because our purpose was to assist in building, testing, and training ML models, we wanted to explore whether the original data and synthetic data produced similar results under the same demographic variables. We have conducted several tests as described below:

**Machine Learning Analyses** The ML findings are presented in Table 4, where we evaluated the efficacy of four different ML models by comparing their performance on both synthetic and original data. We used 80% of the original data to train the model and tested it on 20% of the original data and 100% synthetic data. The outcomes of the four ML models

**Table 2** Descriptive statistics: GATB subtests and performance criterion for original data and synthetic data

|  | Min | Max | Median | Mean | *SD* | *Ku* | *Sk* | Cohen's *d* |
|---|---|---|---|---|---|---|---|---|
| **GATB subtests** | | | | | | | | |
| **G** | | | | | | | | |
|   Original | 69 | 154 | 110 | 109.71 | 14.29 | 2.86 | 0.05 | |
|   Synthetic | 66 | 150 | 110 | 110.10 | 14.93 | 2.49 | 0.01 | − 0.03 |
| **V** | | | | | | | | |
|   Original | 63 | 189 | 109 | 108.94 | 14.73 | 3.87 | 0.22 | |
|   Synthetic | 63 | 166 | 110 | 109.50 | 15.56 | 3.02 | − 0.24 | − 0.04 |
| **N** | | | | | | | | |
|   Original | 63 | 148 | 109 | 109.86 | 13.87 | 3.02 | − 0.16 | |
|   Synthetic | 60 | 146 | 109 | 108.96 | 15.30 | 2.62 | − 0.18 | 0.06 |
| **S** | | | | | | | | |
|   Original | 55 | 163 | 107 | 106.20 | 18.11 | 2.88 | 0.16 | |
|   Synthetic | 53 | 163 | 108 | 107.73 | 18.16 | 2.95 | 0.08 | − 0.08 |
| **P** | | | | | | | | |
|   Original | 68 | 172 | 120 | 119.86 | 17.83 | 2.88 | 0.01 | |
|   Synthetic | 55 | 184 | 121 | 121.16 | 18.33 | 3.03 | − 0.04 | − 0.07 |
| **Q** | | | | | | | | |
|   Original | 80 | 185 | 120 | 120.93 | 16.13 | 3.32 | 0.44 | |
|   Synthetic | 77 | 189 | 120 | 120.95 | 16.84 | 3.20 | 0.40 | − 0.00 |
| **K** | | | | | | | | |
|   Original | 49 | 161 | 115 | 114.33 | 17.08 | 3.06 | − 0.05 | |
|   Synthetic | 59 | 169 | 115 | 114.20 | 17.15 | 3.01 | − 0.01 | 0.01 |
| **F** | | | | | | | | |
|   Original | 43 | 168 | 109 | 108.85 | 19.50 | 2.89 | 0.01 | |
|   Synthetic | 53 | 162 | 107 | 107.02 | 18.41 | 2.75 | 0.03 | 0.10 |
| **M** | | | | | | | | |
|   Original | 57 | 196 | 120 | 118.86 | 20.42 | 3.06 | 0.01 | |
|   Synthetic | 48 | 186 | 115 | 115.26 | 20.62 | 2.79 | 0.12 | 0.18 |
| **Performance Criterion** | | | | | | | | |
| **CRFIN** | | | | | | | | |
|   Original | 39 | 120 | 83 | 83.66 | 12.39 | 3.31 | − 0.12 | |
|   Synthetic | 30 | 118 | 84 | 83.80 | 12.40 | 3.51 | − 0.27 | − 0.01 |

*Note*: *n* = 1056 for both the original dataset and the synthetic dataset. *GATB*, General Aptitude Test Battery. See Table 1 for the names of the GATB measures. *CRFIN*, standard supervisory rating scale; *Ku*, kurtosis; *Sk*, skewness. Cohen's *d* represents the mean difference between the original data and synthetic data

were very similar on both the original test data and the synthetic data, as reflected through metrics such as $R^2$, MSE, and MAE. These findings suggest that the performance of ML models on original data closely mirrors that on synthetic data, indicating the interchangeability of synthetic data.

**Gender and Race Differences in Descriptive Statistics** Gender difference findings are summarized in Table 5. For both males and females, the mean and standard deviation of the synthetic dataset aligned closely with those of the original dataset. Additionally, the Cohen's *d* values were relatively consistent across both datasets. On average, the absolute difference in Cohen's *d* between the synthetic and original datasets for gender was 0.09.

Race difference findings are also presented in Table 5. As with gender, the differences between the White and Black groups in the synthetic data were close to those in the original data. The mean and standard deviation showed similar patterns, and Cohen's *d* maintained directional consistency and largely similar magnitudes. On average, the absolute difference in Cohen's *d* between the synthetic and original datasets for gender was 0.14.

## Study 1 Discussion

This first empirical demonstration showed the application of synthetic data to GATB data. We conducted basic tests and use-specific tests to determine whether synthetic data based on an ML generative model possessed sufficient substitutability

**Table 3** GATB correlation matrices for original and synthetic data

| | G | V | N | S | P | Q | K | F | M | CRFIN |
|---|---|---|---|---|---|---|---|---|---|---|
| **Original:** | | | | | | | | | | |
| G | 1.00 | | | | | | | | | |
| V | 0.78 | 1.00 | | | | | | | | |
| N | 0.76 | 0.51 | 1.00 | | | | | | | |
| S | 0.66 | 0.32 | 0.28 | 1.00 | | | | | | |
| P | 0.45 | 0.31 | 0.34 | 0.49 | 1.00 | | | | | |
| Q | 0.49 | 0.49 | 0.47 | 0.30 | 0.60 | 1.00 | | | | |
| K | 0.20 | 0.21 | 0.22 | 0.10 | 0.37 | 0.36 | 1.00 | | | |
| F | 0.14 | 0.05 | 0.11 | 0.26 | 0.34 | 0.23 | 0.32 | 1.00 | | |
| M | 0.12 | 0.04 | 0.14 | 0.14 | 0.28 | 0.17 | 0.43 | 0.49 | 1.00 | |
| CRFIN | 0.21 | 0.11 | 0.24 | 0.13 | 0.22 | 0.14 | 0.16 | 0.26 | 0.25 | 1.00 |
| **Synthetic:** | | | | | | | | | | |
| G | 1.00 | | | | | | | | | |
| V | 0.74 | 1.00 | | | | | | | | |
| N | 0.75 | 0.57 | 1.00 | | | | | | | |
| S | 0.57 | 0.27 | 0.29 | 1.00 | | | | | | |
| P | 0.50 | 0.33 | 0.43 | 0.59 | 1.00 | | | | | |
| Q | 0.55 | 0.52 | 0.56 | 0.33 | 0.64 | 1.00 | | | | |
| K | 0.23 | 0.23 | 0.33 | 0.10 | 0.40 | 0.42 | 1.00 | | | |
| F | −0.07 | −0.11 | 0.01 | 0.15 | 0.22 | 0.08 | 0.28 | 1.00 | | |
| M | 0.03 | −0.09 | 0.09 | 0.20 | 0.28 | 0.13 | 0.39 | 0.47 | 1.00 | |
| CRFIN | 0.08 | −0.01 | 0.12 | 0.17 | 0.22 | 0.04 | 0.11 | 0.20 | 0.23 | 1.00 |

*Note*: $n = 1056$ for both original data and synthetic data. *GATB*, General Aptitude Test Battery. See Table 1 for the names of the GATB measures. *CRFIN*, standard supervisory rating scale. A correlation magnitude less than 0.05 is not statistically significant

and security. It can be seen that the model estimates were generally quite close. Although we did not show the feedback loop from data evaluation back to data generation in this demonstration, it is important to note that whenever a creator of synthetic data is not satisfied with the evaluation results, they can choose to use different synthetic data generation methods or adjust hyperparameters (see Appendix). This iterative process is important for improving the quality and accuracy of synthetic data. Furthermore, even when the same model and the same hyperparameters are employed, the resulting synthetic data can differ due to the random sampling of data. Differences might be reflected in the value of single data points rather than the overall structure and pattern of the dataset.

We now transition to demonstrate the iterative process with a very different dataset to illustrate the different analytic choices one must consider, depending on the dataset.

## Study 2: Synthetic Data with Multi-source Ratings

For study 2, we applied our synthetic data approach to a multi-source assessment of leadership. Such assessments, also referred to as *360-degree feedback assessments*, are commonly applied in leadership development programs and can help leaders understand how they perceive themselves and how they are perceived by others (Fleenor et al., 2010; Lee & Carpenter, 2018). Areas of consensus, as well as those where disagreements emerge (i.e., "blind spots"), can help inform subsequent leadership development (Atwater et al., 1998).

Because these assessments typically feature multiple rating sources as well as multiple dimensions of leadership, they represent a distinct data structure than what was synthesized in study 1, with a structure that is similar to those found across several areas of the organizational sciences (e.g., assessment centers, performance appraisals; Meriac et al., 2014).

### Participants and Procedure

Study 2 data consist of multi-source ratings of 16,752 leaders. Leaders and their colleagues completed a multi-source leadership assessment that measures dimensions relevant to entry-level leaders. Prior research has found that this assessment exhibits evidence of adequate internal consistency, interrater reliability, content validity, construct validity (e.g., patterns of correlations among dimensions), and criterion-related validity (e.g., correlations with measures of leader effectiveness; Leslie & Braddy, 2015).

**Fig. 3** Correlations between GATB data: histogram of absolute differences between the original and synthetic data. *Note*: The complete correlation difference matrix, featuring differences across 100 cells, is provided in Table 3
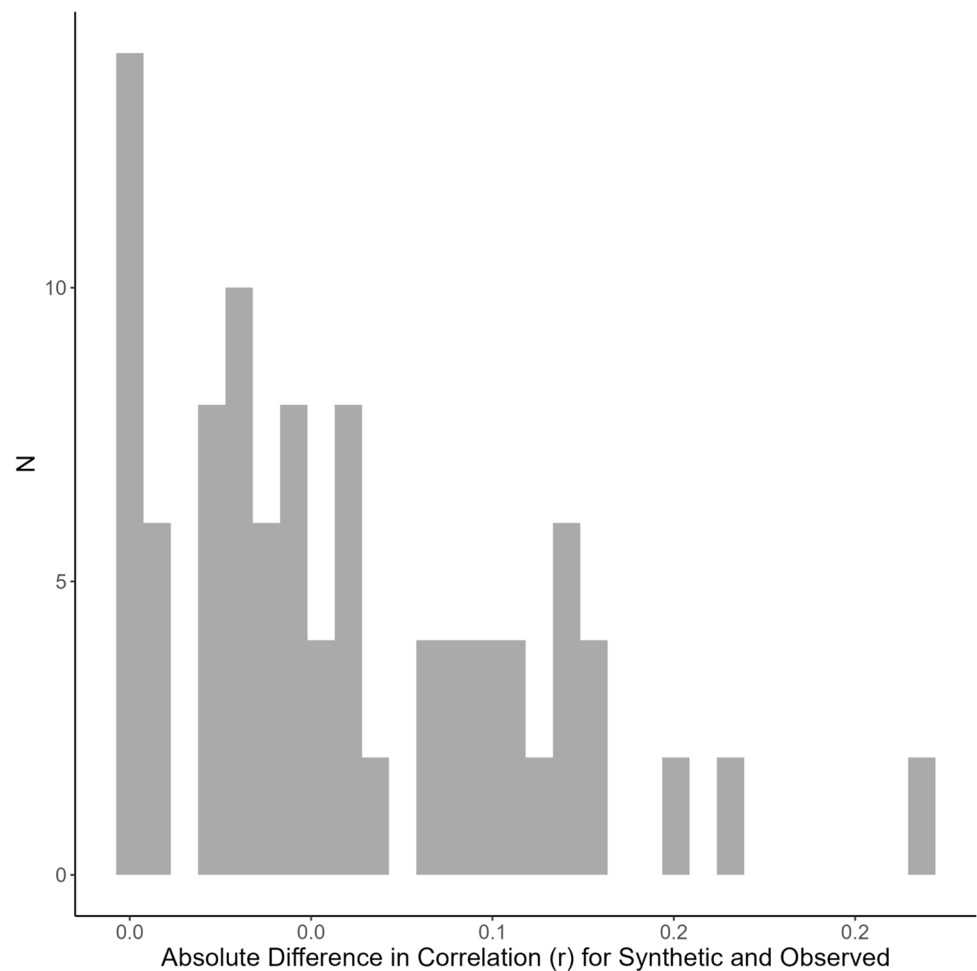


**Table 4** Machine learning analyses comparing original data and synthetic data for supervisory ratings predicted by GATB subtests

|  | XG Boost Regressor | Cat Boost Regressor | Random Forest | LGBM Regressor |
|---|---|---|---|---|
| $R^2$ |  |  |  |  |
| Train original data | 0.17 | 0.18 | 0.17 | 0.20 |
| Test original data | 0.10 | 0.09 | 0.09 | 0.08 |
| Synthetic data | 0.06 | 0.07 | 0.06 | 0.07 |
| MSE |  |  |  |  |
| Train original data | 127.91 | 124.99 | 126.77 | 122.12 |
| Test original data | 137.19 | 139.04 | 138.84 | 139.60 |
| Synthetic data | 144.02 | 142.42 | 144.24 | 142.64 |
| MAE |  |  |  |  |
| Train original data | 8.87 | 8.79 | 8.84 | 8.67 |
| Test original data | 9.17 | 9.27 | 9.26 | 9.32 |
| Synthetic data | 9.55 | 9.49 | 9.54 | 9.47 |

*Note*: $n = 1056$ for both the original dataset and the synthetic dataset. *GATB*, General Aptitude Test Battery; *CRFIN*, standard supervisory rating scale. We used fivefold cross-validation to obtain average values for $R2$, MSE, and MAE. $R^2$, coefficient of determination; *MSE*, mean squared error; *MAE*, mean absolute error

The assessment includes two broad categories that organize the lower-order dimensions: Leading the Organization and Leading Others. These broad categories largely and respectively correspond to the prevailing theoretical models of task- vs. relationship-focused perceptions of leadership behaviors (Gerpott et al., 2019; Meriac et al., 2014; Shaffer et al., 2016). Within the Leading the Organization category, there are four dimensions, while the Leading Others category

**Table 5** GATB subtests and performance criterion standardized mean differences: comparing original and synthetic data by gender and race

| | Male | Female | Cohen's *d* | White | Black | Cohen's *d* |
|---|---|---|---|---|---|---|
| **G** | | | | | | |
| Original | 110.99 (14.51) | 107.99 (13.82) | 0.21 | 111.57 (13.50) | 97.15 (12.76) | 1.10 |
| Synthetic | 111.21 (15.73) | 108.61 (13.67) | 0.18 | 111.74 (14.34) | 98.30 (13.33) | 0.97 |
| **V** | | | | | | |
| Original | 108.19 (15.08) | 109.95 (14.21) | −0.12 | 110.51 (14.08) | 99.52 (13.86) | 0.78 |
| Synthetic | 108.14 (16.85) | 111.31 (13.44) | −0.21 | 110.75 (15.40) | 102.41 (12.45) | 0.60 |
| **N** | | | | | | |
| Original | 111.48 (13.71) | 107.70 (13.79) | 0.27 | 111.17 (13.42) | 100.54 (13.18) | 0.80 |
| Synthetic | 110.34 (15.83) | 107.14 (14.38) | 0.21 | 109.75 (15.35) | 102.31 (13.11) | 0.52 |
| **S** | | | | | | |
| Original | 106.82 (19.22) | 105.38 (16.49) | 0.08 | 107.69 (17.77) | 95.18 (16.91) | 0.72 |
| Synthetic | 108.11 (19.77) | 107.23 (15.76) | 0.05 | 109.36 (17.34) | 95.84 (19.66) | 0.73 |
| **P** | | | | | | |
| Original | 117.59 (18.12) | 122.87 (16.99) | −0.30 | 121.32 (16.95) | 109.65 (19.96) | 0.63 |
| Synthetic | 119.49 (18.76) | 123.39 (17.52) | −0.22 | 123.17 (17.33) | 107.25 (18.83) | 0.88 |
| **Q** | | | | | | |
| Original | 118.12 (15.63) | 124.66 (16.05) | −0.41 | 121.99 (15.83) | 113.40 (16.13) | 0.54 |
| Synthetic | 117.97 (17.50) | 124.92 (15.04) | −0.43 | 122.07 (16.64) | 112.21 (15.72) | 0.61 |
| **K** | | | | | | |
| Original | 112.43 (17.46) | 116.84 (16.24) | −0.26 | 115.12 (16.89) | 108.03 (16.26) | 0.43 |
| Synthetic | 111.30 (17.44) | 118.06 (15.98) | −0.40 | 114.86 (16.91) | 108.03 (16.30) | 0.41 |
| **F** | | | | | | |
| Original | 104.93 (19.39) | 114.08 (18.41) | −0.48 | 109.53 (19.20) | 102.55 (20.27) | 0.35 |
| Synthetic | 101.47 (16.73) | 114.42 (17.97) | −0.75 | 106.77 (18.06) | 106.23 (19.59) | 0.03 |
| **M** | | | | | | |
| Original | 119.40 (21.00) | 118.13 (19.62) | 0.06 | 119.90 (20.21) | 110.45 (20.05) | 0.47 |
| Synthetic | 115.43 (21.59) | 115.04 (19.28) | 0.02 | 115.90 (20.48) | 109.93 (20.96) | 0.29 |
| **CRFIN** | | | | | | |
| Original | 83.65 (12.46) | 83.67 (12.29) | −0.00 | 84.23 (12.24) | 79.19 (12.07) | 0.41 |
| Synthetic | 83.24 (12.22) | 84.56 (12.61) | −0.11 | 84.34 (11.96) | 79.25 (12.84) | 0.41 |

*Note*: *n* = 1056 for both the original dataset and the synthetic dataset. *GATB*, General Aptitude Test Battery. See Table 1 for the names of the GATB measures. Standard deviations are in parentheses. Cohen's *d* represents the difference between male and female effect sizes and the difference between the white and black effect sizes

consists of seven dimensions (see Table 6).[1] The 11 scales were measured using three to thirteen items rated on a 5-point scale ranging from 1 ("To a very little extent") to 5 ("To a very great extent"). Raters also provide ratings of a leader's likelihood to derail, which reflects challenges or issues that if, unaddressed, likely limit one's effectiveness in a leadership position (Atwater et al., 1998).

Aside from the leader's self-ratings, an average of 9.95 raters (*SD* = 3.65) rated each leader. A total of 150,062 raters are reflected in the data, including 65,635 direct reports, 67,507 peers, and 16,920 superiors. For each source, we created aggregate scores for each type of rater and each dimension by averaging across all items within a given type of rater for each dimension. These aggregate scores served as input for our analyses.

## Analyses

To generate synthetic multi-source leadership assessment data, we followed the process described earlier and depicted

---

[1] The assessment also contained dimensions categorized as "Leading Yourself." However, for the sake of parsimony, and because these pertain to constructs beyond typical areas of emphasis within leadership literature (e.g., career management, work-life balance), we have excluded these from our analyses.

**Table 6** Summary of multisource leadership assessment

| Dimension | Sample item |
|---|---|
| *Leading the Organization* | |
| Strategic Perspective | "Links their responsibilities with the mission of the whole organization." |
| Being a Quick Study | "Learns a new skill quickly." |
| Decisiveness | "Is action-oriented." |
| Change Management | "Adapts plans as necessary." |
| *Leading Others* | |
| Leading Employees | "Is willing to delegate important tasks, not just things they don't want to do." |
| Confronting Problem Employees | "Can deal effectively with resistant employees." |
| Participative Management | "Is open to the input of others." |
| Building Collaborative Relationships | "Tries to understand what other people think before making judgments about them." |
| Compassion and Sensitivity | "Is sensitive to signs of overwork in others." |
| Putting People at Ease | "Has personal warmth." |
| Respect for Differences | "Treats people of all backgrounds fairly." |
| *Likelihood of Derailing* | |
| Problems With Interpersonal Relationships | "Is dictatorial in their approach." |
| Difficulty Building and Leading a Team | "Does not resolve conflict among direct reports." |
| Difficulty Changing or Adapting | "Has not adapted to the culture of the organization." |
| Failure To Meet Business Objectives | "Is overwhelmed by complex tasks." |
| Too Narrow a Functional Orientation | "Could not handle management outside of current function." |

*Note*: Dimensions and sample items are drawn from the current multisource measure of leadership (Leslie et al., 2015)

in Fig. 1. First, we examined the original data and found that, much like other rating-based evaluations in the organizational sciences, most variables were significantly skewed and leptokurtic (i.e., non-normal). Thus, most leaders were rated above the midpoint along each competency and below the midpoint for the derailment ratings. Second, we also determined that our intended use of these synthetic data would be to allow others to examine and replicate prior hypothesis testing (Braddy et al., 2014). Specifically, our goal was to synthesize data that reproduced earlier findings, which could then be shared with other researchers interested in self-other agreement leadership research (Fleenor et al., 2010). These steps helped inform our decision to select a CTGAN synthetic data model and to tune several hyperparameters (i.e., epoch and batch size). Although this reflects an iterative process, for the sake of parsimony, we only share the results of the best-performing model (i.e., batch size = 128; epoch = 3000).

## Study 2 Results

### Basic Tests

**Overlapped Sample Test** As with study 1, we began by conducting an overlapped sample test to ensure that the synthetic data did not inadvertently create potential security or privacy risks. This test indicated that there was no overlap

in data cases across the synthetic and original data. This provides more assurance of the security of the synthetic data.

**Constrained Reflection Test** Next, we conducted a constrained reflection test, where we compared the mean and standard deviations for the original and synthetic data. Across the different competencies and measures of derailment, and rating sources, the average absolute Cohen's *d* value was 0.06 (SD = 0.04, min = 0.00, max = 0.19) (see Table 6). This suggests that, on average, the distributions of the variables in the synthetic data exhibited comparable measures of central tendency (i.e., mean) and variability (i.e., standard deviation) when compared to the original data.

**Distribution Kurtosis and Skewness Test** Table 7 also reports the skewness and kurtosis values for each variable in the original and synthetic data. On average, the absolute difference in the skewness and kurtosis values was 0.22 and 0.92. Further inspection revealed that the items pertaining to derailment tended to exhibit the largest differences across the synthetic and original data.

**Variable Correlation Test** We then considered the variable correlation test and examined the differences in the correlations obtained using the synthetic and original data. Across all items and rating sources, we compared 2016 correlations (i.e., each cell in the complete matrix). On average, the absolute difference in the correlation estimates was 0.07

($SD = 0.00$, min $= 0.00$, max $= 0.30$). Figure 4 provides a histogram summarizing these differences, which shows that 80% of the correlation estimates exhibited differences less than or equal to 0.10.

Taken as a whole, we found that the synthetic data exhibited minimal differences in means and standard deviations (i.e., Cohen's $d$) compared to the original data, that skewness and kurtosis were comparable for many variables, and that the majority of correlations were reproduced with relatively small differences. Thus, we proceeded to our use-specific tests.

## Use-Specific Tests

As noted previously, our intended purpose of these synthetic data was to generate data that would reproduce findings previously reported pertaining to the relationship between self-other agreement for task and relationship leadership and derailment (Braddy et al., 2014). Specifically, this model regressed peers' ratings of derailment onto a leader's self-ratings and their superiors' ratings of task and relationship leadership.

To begin, we sought to reduce the number of competencies in the assessment to the two-dimensional space reported in the earlier study (i.e., task and relationship leadership).

Thus, we fit an exploratory factor analysis (EFA) that forced a two-factor solution onto the 11 leadership competencies for both self-ratings and superior's ratings.[2] Across the 22 factor loadings, we found that the average absolute difference between the original and synthetic data was 0.03 ($SD = 0.02$, min $= 0.00$, max $= 0.07$). This suggests that factor analyses returned comparable solutions across the original and synthetic data.

Next, we proceeded to identify more defensible measures of task and relationship leadership by identifying items that contributed to a clear factor structure and were not redundant. Thus, we selected items whose relative loadings (i.e., item loading for a factor/sum of all loadings) were greater than 0.70 for a given. This resulted in two competencies for the task leadership dimension (i.e., decisiveness and confronting problem employees) and two on the relationship leadership dimension (i.e., compassion and putting people at ease). We also calculated a mean derailment score by averaging across peers' ratings on the five derailment items.

---

[2] The EFA was estimated using minimum residuals and varimax rotations (Revelle, 2024). We also inspected a scree plot for the EFA, which indicated that the elbow exhibited a marked flattening for two dimensions. Because each leader was evaluated by multiple raters (e.g., several peers), and we chose to average across multiple raters within a given category, systematic variance in ratings that can be attributable to types of raters are omitted from these models. Thus, these EFA results largely ignore this source of variability in the 360 evaluations.

Using these measures, we estimated polynomial regression models in both the original and synthetic data, which are direct replications of Braddy et al.'s (2014) models. Table 8 reports the model results for both data sets. In general, the two datasets produced models that largely yielded similar results. Specifically, both models indicated that the more a leader and their supervisor rated their task or relationship leadership higher, the less likely the leader's peers were to believe that the leader would derail. Ultimately, this yielded consistent estimates of the two primary response surface tests: the linear effects of the line of congruence (a1) and the linear effects of the line of incongruence (a3). Specifically, for both datasets, the more leaders and their superiors (dis)agreed about their level of task and relationship leadership, the (more) less likely their peers were to indicate that the leader was likely to derail. These findings are largely similar to those reported by Braddy et al. (2014) and, more importantly to our current work, are consistent across both datasets. To allow for further comparisons of the polynomial regression results, we generated response surfaces for each model using the original and synthetic data (see Fig. 5). The surfaces generated when predicting leader derailment using task and relationship ratings are consistent when based on the original and synthetic data.

## Study 2 Discussion

Our second application of synthetic data to an organizational context drew on multi-source leadership ratings. We believe this is a useful addition to study 1 because these data are largely non-normally distributed (i.e., skewed and leptokurtic), reflect a higher dimensional space (i.e., several variables being synthesized at once), and exhibit fairly strong correlations among the measures. Thus, in many ways, we see these data as a relatively stringent test of the capabilities of synthetic data models and suspect that in many other situations (e.g., more normality, less measures, greater discrimination), state-of-the-art synthetic data models would perform even better.

Overall, the synthetic data were able to reproduce previously estimated models using 360 data (Braddy et al., 2014) and produce results that were quite similar to the original data. That said, there were some areas of imprecision. For example, the predictive relationship between self-ratings of task leadership is somewhat larger with the synthetic data than the original data ($-0.20$ vs. $-0.04$), and the non-linear effect of the line of incongruence (i.e., $a_4$ response surface parameters) is non-significant only in the synthetic data for both task and relationship leadership. We suspect that further training of a synthetic data model with additional hyperparameter tuning may reduce these differences. Here, though, it is important to highlight that these comparisons are not based on simulated data.

**Table 7** Descriptive statistics: 360 assessment data and synthetic data

| Dimension | Data | Source | Min | Max | Median | Mean | SD | Kurtosis | Skewness | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| Problems with interpersonal relationships | Original | Superior | 1.00 | 5.00 | 1.25 | 1.44 | 0.58 | 4.99 | 2.02 | |
| | Synthetic | Superior | 1.00 | 5.00 | 1.25 | 1.47 | 0.63 | 5.09 | 2.14 | 0.13 |
| | Original | Direct Report | 1.00 | 5.00 | 1.31 | 1.46 | 0.50 | 4.89 | 1.92 | |
| | Synthetic | Direct Report | 1.00 | 5.00 | 1.37 | 1.52 | 0.58 | 6.17 | 2.25 | 0.08 |
| | Original | Peer | 1.00 | 5.00 | 1.38 | 1.52 | 0.49 | 3.92 | 1.70 | |
| | Synthetic | Peer | 1.00 | 5.00 | 1.36 | 1.47 | 0.45 | 3.24 | 1.56 | 0.07 |
| | Original | Self | 1.00 | 5.00 | 1.25 | 1.43 | 0.47 | 5.35 | 1.80 | |
| | Synthetic | Self | 1.00 | 5.00 | 1.30 | 1.46 | 0.54 | 6.00 | 1.98 | 0.11 |
| Difficulty building and leading a team | Original | Superior | 1.00 | 5.00 | 1.43 | 1.59 | 0.58 | 2.11 | 1.22 | |
| | Synthetic | Superior | 1.00 | 4.62 | 1.52 | 1.58 | 0.56 | 2.66 | 1.38 | 0.08 |
| | Original | Direct Report | 1.00 | 5.00 | 1.43 | 1.55 | 0.48 | 3.32 | 1.47 | |
| | Synthetic | Direct Report | 1.00 | 4.36 | 1.45 | 1.54 | 0.50 | 3.55 | 1.68 | 0.04 |
| | Original | Peer | 1.00 | 5.00 | 1.56 | 1.64 | 0.48 | 2.51 | 1.20 | |
| | Synthetic | Peer | 1.00 | 5.00 | 1.50 | 1.56 | 0.44 | 3.25 | 1.29 | 0.05 |
| | Original | Self | 1.00 | 5.00 | 1.57 | 1.58 | 0.49 | 2.07 | 0.96 | |
| | Synthetic | Self | 1.00 | 3.65 | 1.57 | 1.57 | 0.48 | −0.16 | 0.58 | 0.06 |
| Difficulty changing or adapting | Original | Superior | 1.00 | 5.00 | 1.40 | 1.51 | 0.53 | 3.98 | 1.63 | |
| | Synthetic | Superior | 1.00 | 4.79 | 1.39 | 1.48 | 0.48 | 3.05 | 1.44 | 0.09 |
| | Original | Direct Report | 1.00 | 5.00 | 1.35 | 1.45 | 0.42 | 5.63 | 1.88 | |
| | Synthetic | Direct Report | 1.00 | 4.56 | 1.39 | 1.48 | 0.48 | 8.23 | 2.42 | 0.06 |
| | Original | Peer | 1.00 | 5.00 | 1.48 | 1.56 | 0.43 | 3.95 | 1.48 | |
| | Synthetic | Peer | 1.00 | 4.05 | 1.52 | 1.62 | 0.47 | 2.38 | 1.42 | 0.04 |
| | Original | Self | 1.00 | 5.00 | 1.40 | 1.51 | 0.43 | 4.35 | 1.30 | |
| | Synthetic | Self | 1.00 | 3.57 | 1.42 | 1.50 | 0.42 | 0.80 | 0.94 | 0.07 |
| Failure to meet business objectives | Original | Superior | 1.00 | 5.00 | 1.33 | 1.51 | 0.58 | 3.28 | 1.58 | |
| | Synthetic | Superior | 1.00 | 3.82 | 1.34 | 1.49 | 0.48 | 0.46 | 0.99 | 0.10 |
| | Original | Direct Report | 1.00 | 5.00 | 1.38 | 1.50 | 0.47 | 4.73 | 1.80 | |
| | Synthetic | Direct Report | 1.00 | 5.00 | 1.39 | 1.57 | 0.61 | 7.19 | 2.43 | 0.05 |
| | Original | Peer | 1.00 | 5.00 | 1.50 | 1.60 | 0.47 | 3.06 | 1.39 | |
| | Synthetic | Peer | 1.00 | 4.31 | 1.48 | 1.57 | 0.49 | 2.78 | 1.44 | 0.12 |
| | Original | Self | 1.00 | 5.00 | 1.50 | 1.59 | 0.49 | 2.05 | 1.02 | |
| | Synthetic | Self | 1.00 | 3.72 | 1.50 | 1.59 | 0.50 | 0.23 | 0.81 | 0.11 |
| Too narrow a functional orientation | Original | Superior | 1.00 | 5.00 | 1.70 | 1.82 | 0.74 | 0.72 | 0.98 | |
| | Synthetic | Superior | 1.00 | 5.00 | 1.81 | 1.84 | 0.71 | 0.39 | 0.81 | 0.07 |
| | Original | Direct Report | 1.00 | 5.00 | 1.47 | 1.59 | 0.51 | 3.42 | 1.54 | |
| | Synthetic | Direct Report | 1.00 | 5.00 | 1.48 | 1.58 | 0.56 | 6.89 | 2.22 | 0.01 |
| | Original | Peer | 1.00 | 5.00 | 1.73 | 1.83 | 0.58 | 1.21 | 0.97 | |
| | Synthetic | Peer | 1.00 | 4.95 | 1.77 | 1.87 | 0.65 | 0.58 | 0.88 | 0.02 |
| | Original | Self | 1.00 | 5.00 | 1.60 | 1.72 | 0.58 | 0.92 | 0.84 | |
| | Synthetic | Self | 1.00 | 4.04 | 1.78 | 1.73 | 0.56 | −0.31 | 0.53 | 0.17 |
| Strategic perspective | Original | Superior | 1.13 | 5.00 | 4.13 | 4.09 | 0.54 | 0.60 | −0.53 | |
| | Synthetic | Superior | 1.13 | 5.00 | 4.09 | 4.05 | 0.56 | 2.12 | −1.01 | 0.03 |
| | Original | Direct Report | 1.54 | 5.00 | 4.36 | 4.30 | 0.41 | 2.22 | −1.06 | |
| | Synthetic | Direct Report | 1.54 | 5.00 | 4.37 | 4.35 | 0.41 | 2.94 | −1.14 | 0.07 |
| | Original | Peer | 1.33 | 5.00 | 4.20 | 4.16 | 0.41 | 1.53 | −0.75 | |
| | Synthetic | Peer | 1.33 | 5.00 | 4.20 | 4.12 | 0.51 | 2.37 | −1.19 | 0.07 |
| | Original | Self | 1.50 | 5.00 | 4.00 | 4.04 | 0.44 | 0.07 | −0.12 | |
| | Synthetic | Self | 2.25 | 5.00 | 4.01 | 4.07 | 0.45 | −0.43 | 0.10 | 0.12 |
| Being a quick study | Original | Superior | 1.00 | 5.00 | 4.33 | 4.23 | 0.59 | 0.52 | −0.61 | |
| | Synthetic | Superior | 2.54 | 5.00 | 4.24 | 4.25 | 0.55 | −0.40 | −0.40 | 0.03 |

**Table 7** (continued)

| Dimension | Data | Source | Min | Max | Median | Mean | SD | Kurtosis | Skewness | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | Original | Direct Report | 1.33 | 5.00 | 4.33 | 4.30 | 0.47 | 2.29 | −1.08 | |
| | Synthetic | Direct Report | 1.33 | 5.00 | 4.33 | 4.32 | 0.51 | 5.08 | −1.65 | 0.04 |
| | Original | Peer | 1.00 | 5.00 | 4.25 | 4.21 | 0.45 | 1.72 | −0.82 | |
| | Synthetic | Peer | 1.78 | 5.00 | 4.24 | 4.23 | 0.45 | 1.52 | −0.81 | 0.13 |
| | Original | Self | 1.33 | 5.00 | 4.00 | 3.95 | 0.60 | −0.31 | −0.14 | |
| | Synthetic | Self | 2.02 | 5.00 | 4.00 | 3.95 | 0.60 | −0.25 | −0.15 | 0.05 |
| Decisiveness | Original | Superior | 1.00 | 5.00 | 4.00 | 4.07 | 0.63 | 0.61 | −0.62 | |
| | Synthetic | Superior | 1.51 | 5.00 | 4.03 | 4.08 | 0.60 | 0.56 | −0.57 | 0.00 |
| | Original | Direct Report | 1.22 | 5.00 | 4.33 | 4.24 | 0.47 | 1.88 | −0.98 | |
| | Synthetic | Direct Report | 1.22 | 5.00 | 4.26 | 4.24 | 0.47 | 4.05 | −1.35 | 0.02 |
| | Original | Peer | 1.33 | 5.00 | 4.17 | 4.11 | 0.47 | 1.11 | −0.69 | |
| | Synthetic | Peer | 1.33 | 5.00 | 4.16 | 4.10 | 0.51 | 1.33 | −0.81 | 0.01 |
| | Original | Self | 1.00 | 5.00 | 4.00 | 3.90 | 0.61 | 0.05 | −0.27 | |
| | Synthetic | Self | 1.98 | 5.00 | 4.00 | 3.89 | 0.61 | −0.02 | −0.39 | 0.06 |
| Change management | Original | Superior | 1.00 | 5.00 | 4.00 | 4.00 | 0.53 | 0.57 | −0.33 | |
| | Synthetic | Superior | 1.42 | 5.00 | 4.00 | 4.04 | 0.47 | 1.06 | −0.21 | 0.04 |
| | Original | Direct Report | 1.44 | 5.00 | 4.20 | 4.17 | 0.45 | 1.48 | −0.81 | |
| | Synthetic | Direct Report | 1.44 | 5.00 | 4.18 | 4.13 | 0.49 | 3.61 | −1.42 | 0.07 |
| | Original | Peer | 1.00 | 5.00 | 4.07 | 4.05 | 0.43 | 1.27 | −0.57 | |
| | Synthetic | Peer | 1.79 | 5.00 | 4.04 | 3.98 | 0.53 | 1.26 | −0.81 | 0.12 |
| | Original | Self | 1.78 | 5.00 | 3.89 | 3.92 | 0.45 | 0.05 | 0.09 | |
| | Synthetic | Self | 2.51 | 5.00 | 3.98 | 3.94 | 0.48 | −0.17 | 0.10 | 0.09 |
| Leading employees | Original | Superior | 1.00 | 5.00 | 4.00 | 3.98 | 0.53 | 0.45 | −0.35 | |
| | Synthetic | Superior | 1.51 | 5.00 | 4.00 | 4.01 | 0.55 | 0.66 | −0.38 | 0.06 |
| | Original | Direct Report | 1.09 | 5.00 | 4.18 | 4.13 | 0.48 | 1.44 | −0.86 | |
| | Synthetic | Direct Report | 1.09 | 5.00 | 4.17 | 4.17 | 0.51 | 1.82 | −0.94 | 0.04 |
| | Original | Peer | 1.25 | 5.00 | 4.05 | 4.02 | 0.45 | 1.16 | −0.60 | |
| | Synthetic | Peer | 1.25 | 5.00 | 3.98 | 3.93 | 0.53 | 4.11 | −1.35 | 0.05 |
| | Original | Self | 1.69 | 5.00 | 3.85 | 3.89 | 0.45 | −0.03 | 0.02 | |
| | Synthetic | Self | 2.18 | 5.00 | 3.85 | 3.86 | 0.48 | −0.20 | 0.03 | 0.03 |
| Confronting problem employees | Original | Superior | 1.00 | 5.00 | 3.83 | 3.74 | 0.65 | 0.26 | −0.39 | |
| | Synthetic | Superior | 1.21 | 5.00 | 3.75 | 3.75 | 0.59 | 0.54 | −0.33 | 0.00 |
| | Original | Direct Report | 1.00 | 5.00 | 4.00 | 3.96 | 0.56 | 0.93 | −0.71 | |
| | Synthetic | Direct Report | 1.00 | 5.00 | 4.00 | 4.01 | 0.57 | 1.61 | −0.92 | 0.02 |
| | Original | Peer | 1.00 | 5.00 | 3.89 | 3.84 | 0.56 | 0.90 | −0.61 | |
| | Synthetic | Peer | 1.00 | 5.00 | 3.86 | 3.86 | 0.61 | 1.03 | −0.71 | 0.00 |
| | Original | Self | 1.25 | 5.00 | 3.50 | 3.57 | 0.58 | 0.11 | −0.13 | |
| | Synthetic | Self | 1.25 | 5.00 | 3.58 | 3.56 | 0.59 | 0.21 | −0.25 | 0.01 |
| Participative management | Original | Superior | 1.00 | 5.00 | 4.00 | 4.08 | 0.55 | 0.59 | −0.45 | |
| | Synthetic | Superior | 1.40 | 5.00 | 4.06 | 4.05 | 0.54 | 1.40 | −0.78 | 0.02 |
| | Original | Direct Report | 1.56 | 5.00 | 4.22 | 4.17 | 0.48 | 1.57 | −0.92 | |
| | Synthetic | Direct Report | 1.56 | 5.00 | 4.25 | 4.21 | 0.52 | 2.44 | −1.17 | 0.07 |
| | Original | Peer | 1.00 | 5.00 | 4.11 | 4.07 | 0.46 | 1.43 | −0.75 | |
| | Synthetic | Peer | 1.17 | 5.00 | 4.12 | 4.04 | 0.56 | 1.88 | −1.04 | 0.08 |
| | Original | Self | 1.89 | 5.00 | 4.00 | 4.01 | 0.46 | −0.06 | −0.05 | |
| | Synthetic | Self | 2.14 | 5.00 | 4.00 | 3.99 | 0.46 | −0.09 | −0.11 | 0.14 |
| Building collaborative relationships | Original | Superior | 1.00 | 5.00 | 4.00 | 4.03 | 0.59 | 0.70 | −0.59 | |
| | Synthetic | Superior | 1.14 | 5.00 | 4.02 | 4.03 | 0.58 | 1.01 | −0.64 | 0.03 |
| | Original | Direct Report | 1.43 | 5.00 | 4.24 | 4.17 | 0.49 | 1.82 | −1.00 | |
| | Synthetic | Direct Report | 1.43 | 5.00 | 4.19 | 4.13 | 0.56 | 3.03 | −1.38 | 0.05 |

**Table 7** (continued)

| Dimension | Data | Source | Min | Max | Median | Mean | SD | Kurtosis | Skewness | Cohen's $d$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | Original | Peer | 1.00 | 5.00 | 4.10 | 4.05 | 0.49 | 1.43 | −0.84 | |
| | Synthetic | Peer | 1.00 | 5.00 | 4.15 | 4.11 | 0.54 | 1.63 | −0.87 | 0.09 |
| | Original | Self | 1.29 | 5.00 | 4.00 | 3.94 | 0.45 | 0.11 | −0.03 | |
| | Synthetic | Self | 2.72 | 5.00 | 4.00 | 3.96 | 0.47 | −0.33 | 0.09 | 0.19 |
| Compassion and sensitivity | Original | Superior | 1.00 | 5.00 | 4.17 | 4.20 | 0.50 | 0.36 | −0.44 | |
| | Synthetic | Superior | 2.03 | 5.00 | 4.20 | 4.19 | 0.49 | 0.39 | −0.53 | 0.06 |
| | Original | Direct Report | 1.60 | 5.00 | 4.31 | 4.25 | 0.45 | 1.52 | −0.92 | |
| | Synthetic | Direct Report | 1.60 | 5.00 | 4.31 | 4.29 | 0.46 | 2.23 | −1.08 | 0.02 |
| | Original | Peer | 1.00 | 5.00 | 4.18 | 4.16 | 0.44 | 1.70 | −0.75 | |
| | Synthetic | Peer | 1.00 | 5.00 | 4.14 | 4.08 | 0.48 | 2.57 | −1.09 | 0.08 |
| | Original | Self | 1.33 | 5.00 | 4.00 | 4.05 | 0.49 | 0.10 | −0.30 | |
| | Synthetic | Self | 2.56 | 5.00 | 4.01 | 4.10 | 0.50 | −0.88 | 0.10 | 0.03 |
| Putting people at ease | Original | Superior | 1.00 | 5.00 | 4.33 | 4.34 | 0.60 | 0.57 | −0.83 | |
| | Synthetic | Superior | 2.37 | 5.00 | 4.34 | 4.36 | 0.55 | −0.03 | −0.62 | 0.02 |
| | Original | Direct Report | 1.00 | 5.00 | 4.44 | 4.36 | 0.50 | 1.79 | −1.16 | |
| | Synthetic | Direct Report | 1.72 | 5.00 | 4.48 | 4.39 | 0.51 | 2.43 | −1.36 | 0.05 |
| | Original | Peer | 1.00 | 5.00 | 4.39 | 4.32 | 0.49 | 1.57 | −0.99 | |
| | Synthetic | Peer | 1.40 | 5.00 | 4.43 | 4.36 | 0.50 | 0.87 | −0.90 | 0.08 |
| | Original | Self | 1.00 | 5.00 | 4.00 | 4.05 | 0.61 | −0.26 | −0.30 | |
| | Synthetic | Self | 1.87 | 5.00 | 4.01 | 4.10 | 0.64 | −0.23 | −0.44 | 0.06 |
| Respect for differences | Original | Superior | 1.00 | 5.00 | 4.50 | 4.49 | 0.51 | 1.09 | −0.93 | |
| | Synthetic | Superior | 2.74 | 5.00 | 4.51 | 4.52 | 0.46 | 0.41 | −0.89 | 0.05 |
| | Original | Direct Report | 1.25 | 5.00 | 4.58 | 4.52 | 0.39 | 3.95 | −1.43 | |
| | Synthetic | Direct Report | 1.62 | 5.00 | 4.64 | 4.56 | 0.38 | 5.50 | −1.70 | 0.01 |
| | Original | Peer | 1.00 | 5.00 | 4.50 | 4.45 | 0.39 | 2.91 | −1.06 | |
| | Synthetic | Peer | 2.17 | 5.00 | 4.49 | 4.41 | 0.41 | 1.66 | −1.02 | 0.09 |
| | Original | Self | 2.00 | 5.00 | 4.50 | 4.44 | 0.50 | −0.18 | −0.63 | |
| | Synthetic | Self | 3.04 | 5.00 | 4.74 | 4.50 | 0.49 | −0.75 | −0.60 | 0.12 |

*Note*: $n = 16{,}752$ for the original dataset and 15,000 for the synthetic dataset. Cohen's $d$ represents the difference between the original data and synthetic data effect sizes

That is, we generated the data synthetically and then began conducting analyses and fitting potential models. Thus, the data are not geared towards a single analysis (e.g., exploratory factor analysis, polynomial regression model), but instead are intended to preserve many of the features of the original data.

Lastly, it is important to reemphasize that developing a suitable synthetic data generation model is usually an iterative process. Specifically, as indicated in Fig. 1, we examined several alternative synthetic data generation models (e.g., Copula GAN) and employed a range of hyperparameters (e.g., batch sizes that included 128, 256, and 500 and epochs including 500, 1000, 2000, and 3000).[3] For the sake of parsimony, though, we only share the results of the

best-performing model (i.e., batch size = 128; epoch = 3000). However, it is critical to highlight that researchers who generate synthetic data will likely need to consider several data generation models and employ a range of hyperparameters until they are satisfied with the results of the various quality tests.

## General Discussion

In this research, we propose a systematic framework to guide the generation and evaluation of synthetic data with the hopes of facilitating the application of synthetic data in organizational research. Using this framework, we present two empirical demonstrations in studies 1 and 2 to illustrate how synthetic data can serve as a potential solution when providing the original data is impractical, restricted, or poses privacy concerns. Our goal is to provide organizational

---

[3] Complete results of the quality tests for these alternative synthetic data models are available on GitHub: https://github.com/wpengda/SyntheticData_OrganizationalScience.

**Fig. 4** Correlations between leadership assessments: histogram of absolute differences between the original and synthetic data. *Note*: The complete correlation difference matrix, featuring differences across 2080 cells, is provided in our GitHub repository
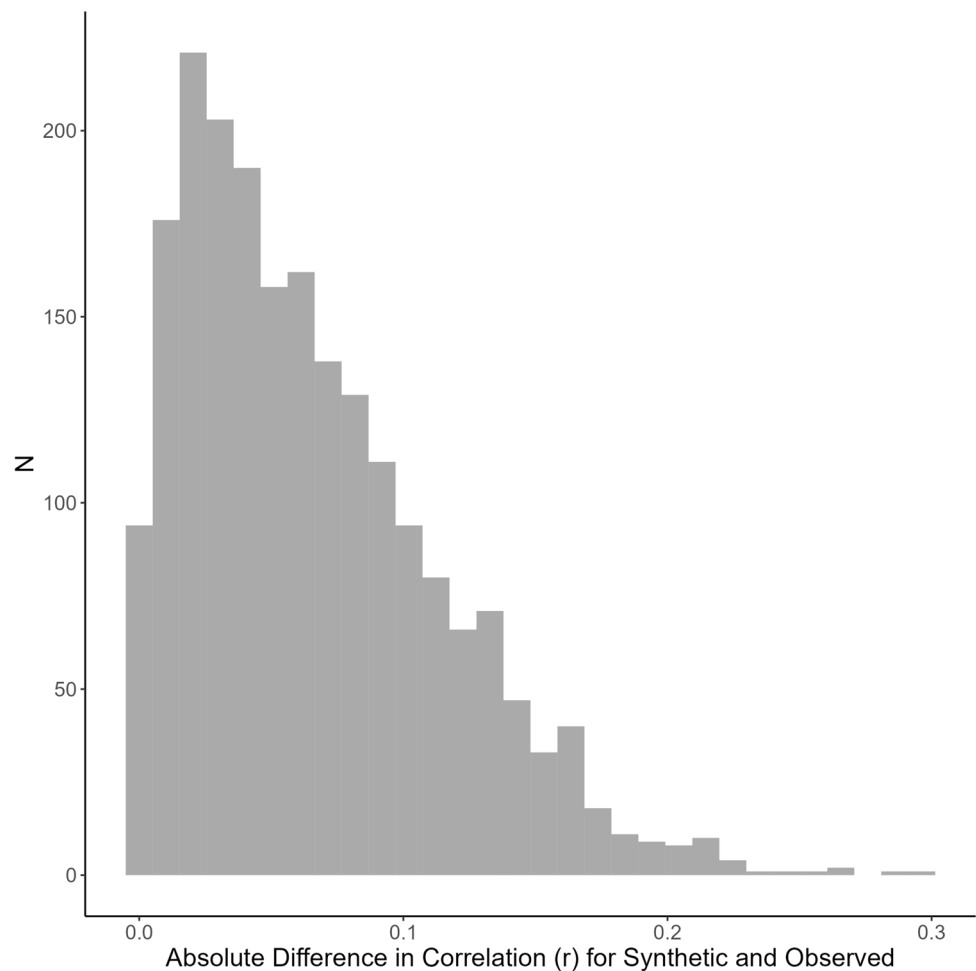


researchers and practitioners with a general understanding of the principles behind generating synthetic data, inspiring them to further explore synthetic data methodologies and their use in organizational research.

## Potential Limitations and Ethical Considerations

The use of synthetic data offers a realm of fresh prospects for exploration and advancement. However, like every new advancement, it comes with particular challenges. We want to emphasize that the use of synthetic data is best-suited as an alternative when sharing original data is impractical, restricted, or poses privacy concerns. Synthetic data are inherently a variation of the original data. Therefore, subsequent models or analyses performed using synthetic data can approach, and often very closely approximate, the original data, there will still be variations and perturbations. We believe that the recommended framework and the quality-checking methods proposed we have proposed can help users of synthetic data effectively assess the quality of the generated data and ensure it meets their intended purposes. Along with leveraging our proposed framework, we would

also recommend that researchers, whenever possible, make their original data available.

If sharing the original data is not possible, and researchers intend to synthesize data, it is important to note that generating these data is not a standardized or "plug-and-play" process but rather one that requires careful attention, involves multiple decision points, and should be transparently documented and described to the eventual end-user. Again, we would like to emphasize the importance of the "Use-Specific Tests" in our proposed process, which should be closely tailored to the intended use of synthetic data. For example, study 1 featured measures of cognitive ability (i.e., GATB), so ensuring comparable subgroup differences was a critical use-specific test. In contrast, study 2 featured multisource ratings of leadership, making the similarity in response surface analyses more relevant.

We also believe that synthetic data, although allowing for greater data sharing and science-practice partnerships, requires further consideration and collective experience among organizational scientists about how synthetic data should be generated and what arrangements might work best. For example, establishing mechanisms by which an

**Table 8** Self-other agreement polynomial regression models for original and synthetic data

| | Perceptions of task leadership | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| | Original data | | | Synthetic data | | |
| Variable | $b$ | SE | $p$ | $b$ | SE | $p$ |
| Self-ratings | $-0.04$ | 0.06 | 0.55 | $-0.20$ | 0.07 | $<0.01$ |
| Supervisor ratings | $-0.36$ | 0.06 | $<0.01$ | $-0.64$ | 0.06 | $<0.01$ |
| Self-ratings$^2$ | 0.00 | 0.01 | 0.94 | 0.00 | 0.01 | 0.68 |
| Self-ratings * supervisor ratings | 0.03 | 0.01 | 0.02 | 0.07 | 0.01 | $<0.01$ |
| Supervisor ratings$^2$ | 0.02 | 0.01 | 0.01 | 0.03 | 0.01 | $<0.01$ |
| Response surface tests | | | | | | |
| $a_1$ | $-0.40$ | 0.09 | $<0.01$ | $-0.85$ | 0.10 | $<0.01$ |
| $a_2$ | 0.05 | 0.01 | $<0.01$ | 0.10 | 0.01 | $<0.01$ |
| $a_3$ | 0.32 | 0.08 | $<0.01$ | 0.44 | 0.09 | $<0.01$ |
| $a_4$ | $-0.01$ | 0.02 | 0.66 | $-0.04$ | 0.02 | 0.04 |
| | Perceptions of relationship leadership | | | | | |
| | Original data | | | Synthetic data | | |
| Variable | $B$ | SE | $p$ | $b$ | SE | $p$ |
| Self-ratings | $-0.07$ | 0.09 | 0.44 | $-0.06$ | 0.10 | 0.53 |
| Supervisor ratings | $-0.67$ | 0.09 | $<0.01$ | $-1.03$ | 0.10 | $<0.01$ |
| Self-ratings$^2$ | $-0.03$ | 0.01 | 0.02 | $-0.05$ | 0.01 | $<0.01$ |
| Self-ratings * supervisor ratings | 0.04 | 0.02 | $<0.01$ | 0.09 | 0.02 | $<0.01$ |
| Supervisor ratings$^2$ | 0.04 | 0.01 | $<0.01$ | 0.06 | 0.01 | $<0.01$ |
| Response surface tests | | | | | | |
| $a_1$ | $-0.74$ | 0.14 | $<0.01$ | $-1.09$ | 0.15 | $<0.01$ |
| $a_2$ | 0.05 | 0.02 | $<0.01$ | 0.09 | 0.02 | $<0.01$ |
| $a_3$ | 0.60 | 0.12 | $<0.01$ | 0.97 | 0.13 | $<0.01$ |
| $a_4$ | $-0.03$ | 0.03 | 0.18 | $-0.08$ | 0.03 | $<0.01$ |

*Note*: $n = 16{,}752$ for the original dataset and 15,000 for the synthetic dataset. $R^2$ for ratings of derailment based on the polynomial regression models was 0.12 when using the original data and 0.10 for the synthetic data

organization could verify synthetic data findings with the original data that it possesses. Also, transparent reporting of the synthetic data generation process, as discussed in this paper, is critical. For example, researchers should clearly state from the outset that they are using synthetic data. They should describe how the synthetic data were generated, including details of the data generation models and hyperparameters used, and disclose challenges encountered during the synthesis process (see Appendix for more information about hyperparameters).

In addition, the quality of synthetic data is highly dependent on the quality of the original data used to generate it. If the original data contains errors, the synthetic counterpart will also reflect those flaws. For example, if the nature of the variables is misspecified in the original data, then this will be recreated in the synthetic data. Likewise, if the original data sample size is a relatively small sample, then the usefulness of synthetic data may become limited. Importantly, this consideration applies to the intended level of analysis and theory (e.g., individuals, teams, leaders, or organizations) rather than the level at which the data were collected.

Furthermore, generating high-quality synthetic depends on the quality of the generative model used. If the model does not properly capture the key characteristics of the original data, the generated data may not provide accurate insights. In our framework, we refrained from advancing uniform standards for evaluating synthetic data. This was an intentional decision because we believe that doing so would be more harmful than beneficial. Such thresholds could impose unnecessary restrictions on the practical application of synthetic data, potentially misleading researchers to prioritize a single "compliance" goal rather than carefully considering the applicability and efficaciousness of the data in their specific contexts (Kreamer et al., 2023). Overly strict thresholds may also encourage a narrow approach, limiting flexibility and creativity, and preventing researchers from fully leveraging the potential of synthetic data in complex, dynamic real-world applications. We advocate for a context-specific evaluation framework, allowing researchers to adjust their evaluation methods based on different application scenarios, while also emphasizing the importance of transparency in documenting these details. In fact, for some
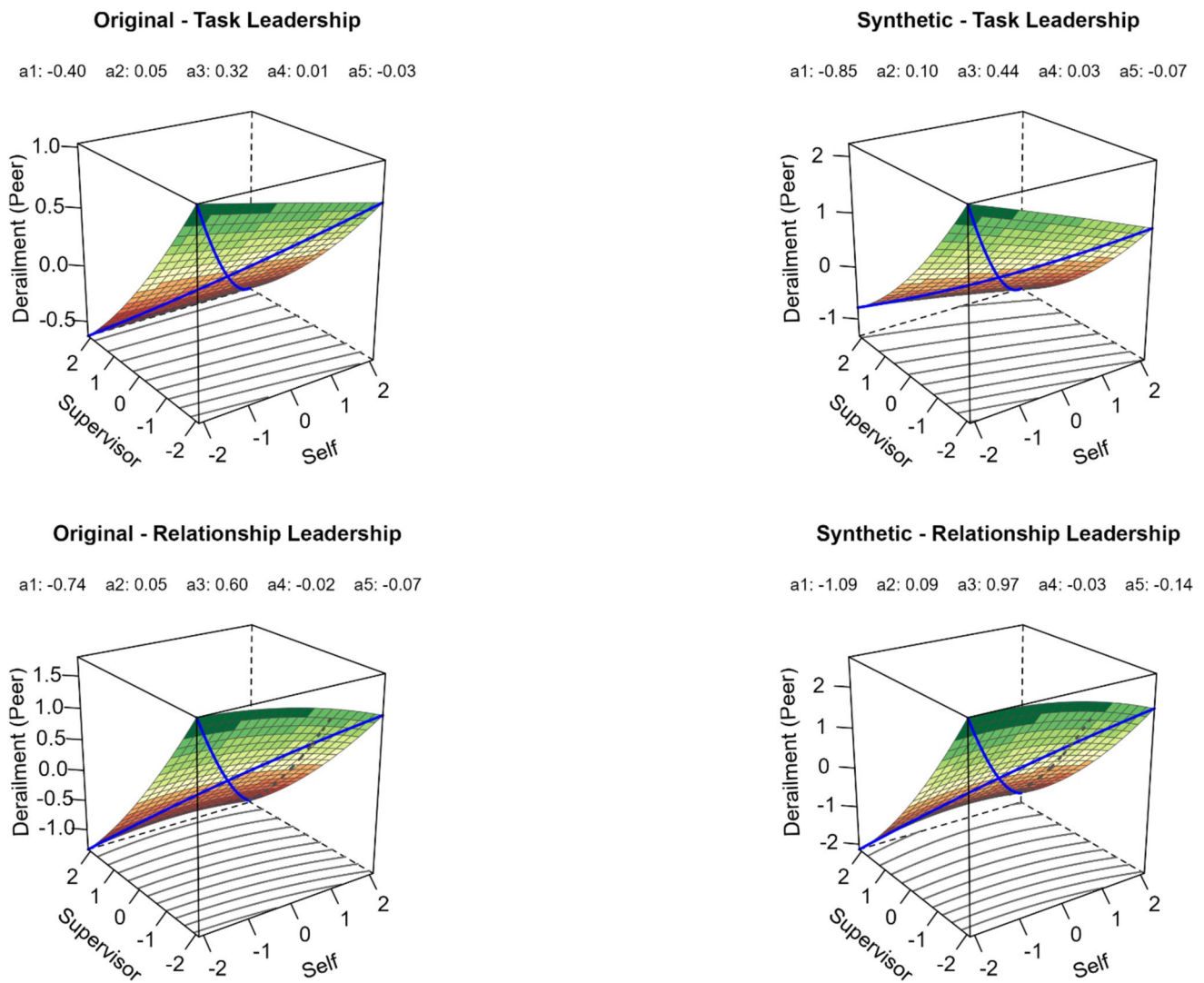
**Original - Task Leadership**

a1: -0.40    a2: 0.05    a3: 0.32    a4: 0.01    a5: -0.03



**Synthetic - Task Leadership**

a1: -0.85    a2: 0.10    a3: 0.44    a4: 0.03    a5: -0.07



**Original - Relationship Leadership**

a1: -0.74    a2: 0.05    a3: 0.60    a4: -0.02    a5: -0.07



**Synthetic - Relationship Leadership**

a1: -1.09    a2: 0.09    a3: 0.97    a4: -0.03    a5: -0.14



**Fig. 5** Response surfaces for polynomial regression models based on original and synthetic data. *Note*: Responses surfaces are based on the results of the polynomial regression models reported in Table 8. Surfaces on the left-hand side correspond to those based on models using the original data, while the surfaces on the right-hand side are based on models using the synthetic data. $n = 16,752$ for the original dataset and 15,000 for the synthetic data

specific applications or complex data structures, further research and development may be required to generate synthetic data effectively. Thus, it might behoove scholars interested in synthesizing data to recognize that this approach requires some technical and algorithmic knowledge. Ideally, the materials we have shared in our online repository can help those interested begin to increase their knowledge of this approach.

## Data Sharing Recommendations

As mentioned above, synthetic data provides a promising solution for improving the openness of our research. Nevertheless, synthetic data will not perfectly replicate the original data, as it is inherently a variant of these data. Therefore,

problems that can be solved by providing partial information from the original data while maintaining privacy may be preferable to using synthetic data. For example, providing aggregated statistical results from the original data—which may be less likely to compromise privacy—could be a potential solution. If this is not feasible, synthetic data seems like a useful alternative.

Additionally, we encourage researchers to ensure that they abide by prevailing ethical norms, specific rules, and organizational regulations regarding data sharing. When in doubt regarding data sharing, the authors should contact stakeholders who can provide oversight and guidance on such practices. For instance, Institutional Review Boards may recommend an explicit statement within the informed consent forms for human subjects indicating that synthetic

data may be generated in lieu of the original data (e.g., "For future research studies, we may use the data you provide to generate de-identified synthetic data. These non-identifiable, synthetic data may be shared with other researchers for future research studies without requiring additional consent from you."). Other entities, such as legal departments, data governance committees, or database owners, may have additional recommendations and guidance that pertain to the use of synthetic data.

### Future Research

We believe that synthetic data offers organizational scholars an opportunity to use previously inaccessible data. Given its novelty, future organizational research exploring the application of synthetic data would be beneficial.

For example, future research may consider the processes for, and benefits of, generating multiple synthetic datasets from a single original dataset. Across these different datasets, one could emphasize different criteria that might be relevant for various evaluations. Due to the inherent randomness of ML methods, each synthetic dataset generated from an original dataset is bound to be different. By comprehensive analysis of the set of synthetic datasets, future research could shed further light on the nature of the original dataset and the robustness of the synthetic dataset generation.

Aside from this, the main focus of this paper is the discussion and examination of numeric data. However, data types are not limited to numeric data; other types, such as textual data, also exist. Such text-based data often emerge via qualitative methodologies (e.g., interviews or focus group transcripts). Rapidly advancing large language models (LLMs) may be a good choice for generating synthetic data in the form of textual data (e.g., Li et al. 2023; Tang et al. 2023). Furthermore, recent studies have shown that LLMs exhibit remarkable abilities to produce human-like responses across various psychological tasks (e.g., Caron and Srivastava, 2022; Huang et al. 2023; Jiang et al. 2023; Lampinen et al. 2024; Pellert et al. 2024; Wang et al. 2024a). Based on this foundation, agent-based simulations to generate synthetic data are also worth exploring (e.g., Chan et al. 2024; Chen et al. 2024; Wang et al., 2024b).

Moreover, we think encouraging cross-disciplinary collaborations—such as facilitating exchanges between computer scientists, statisticians, and organizational scientists—can introduce unprecedented methods and ideas. Through closer cooperation among multiple disciplines, we can fully explore and utilize the vast potential of synthetic data.

## Conclusion

The open science movement has sought to promote the sharing of data as such a practice can serve to accelerate the advancement of theory, practice, and policy making. However, concerns remain about the need to protect human subjects and proprietary information. The creation of synthetic data is one means to promote open data while mitigating potential ethical and legal concerns. In the current paper, we reviewed the key steps in an iterative process needed to generate synthetic data. These include the consideration of a number of assumptions as well as basic and specific-use tests in order to evaluate the quality of synthetic data. We then demonstrated the use of these steps on two unique example datasets. We hope that the synthetic data approach, as illustrated here across two studies, serves to promote greater sharing of data in the organizational sciences.

## Appendix

### Parameters and Hyperparameters

Before we begin, it is important to define some terms commonly used in the field of ML, similar to how various terms are defined in psychology. ML is gaining increasing attention in the field of personnel selection (Koenig et al., 2023), but people may still not be familiar with the related terminology. We will introduce the concepts of parameters and hyperparameters in the sections below.

*Parameters* are elements that can be initialized and updated through the learning process, such as the weights of neurons in a neural network. In contrast, *hyperparameters*, which cannot be directly estimated from data, must be set before training the ML model. They also define the model's architecture (Kuhn & Johnson, 2013). The process of designing an ideal model architecture with the best configuration of hyperparameters is known as hyperparameter optimization/tuning. This tuning is considered an important component in building effective ML models (Hutter et al., 2019).

### Common Hyperparameters

Common hyperparameters include as follows (Hutter et al., 2019; Yang & Shami, 2020):

1. **Learning rates:** These rates determine the speed at which network parameters are updated during training, directly impacting learning efficiency.
2. **Number of epochs:** This defines how many times the complete dataset is used to train the model, which can affect both training duration and model accuracy.

3. **Batch size:** This hyperparameter sets the amount of data fed into the model during each training iteration, affecting both the speed and stability of the training process.

4. **Optimization algorithms:** Optimization algorithm refers to a method employed in ML to adjust the parameters of a model. The goal is to either minimize or maximize a specific function, often the loss function. Options such as Adam or SGD can significantly influence the effectiveness and speed of training.

5. **Network architecture:** Elements like the number of hidden layers and neurons per layer play important roles in determining the model's learning capability and complexity.

6. **Noise dimension:** The dimension of the input noise vector for the Generator influences the diversity of the data generated.

## Loss Function and Optimize Hyperparameters

In addition, we would like to briefly introduce how to optimize hyperparameters. Optimizing hyperparameters is usually aimed at minimizing the loss function. The loss function is a measure of the difference between the model's output and the true labels; for synthetic data, it measures how close the output is to the original data. By minimizing this function, the synthetic data generated by the model can be improved. Generally speaking, different goals and models will choose different loss functions. Common loss functions include the MSE and the MAE. GAN-based models include the generator and the discriminator, each with its own loss function. The following are common methods for optimizing hyperparameters (Yang & Shami, 2020).

1. **Model-free algorithms optimization:** These refer to optimization methods that do not rely on a clear mathematical description of the underlying model. For example, Manual Search is a very basic method where hyperparameters are manually adjusted by individuals based on their experience and intuition, also known as "trial and error" or babysitting (Abreu, 2019). Another example is Grid Search, which is essentially a method of exhaustive search. For each hyperparameter, users select a small finite set to explore (Goodfellow et al., 2016). The Cartesian product of these hyperparameters results in several combinations, and Grid Search trains models using each combination to select the one with the smallest loss function value as the best hyperparameter (Hutter et al., 2019).

2. **Gradient-based optimization:** This is a traditional optimization technique that computes the gradient of variables to identify promising directions and move towards the optimal solution (Bengio, 2000). By randomly selecting a data point or a small subset of data, this technique updates the model parameters by moving in the opposite direction of the gradient computed for that sample or subset, thus taking a step towards minimizing the loss function. Therefore, after convergence, it can achieve a local optimum. For certain ML algorithms, the gradient of some hyperparameters can be computed, and then gradient descent is used to optimize these hyperparameters. However, it can only be used to optimize continuous hyperparameters, as other types of hyperparameters (e.g., the depth of the decision tree; the number of layers in the network structure; and the choice of activation function) do not have a gradient direction. Moreover, this is only effective for convex functions, as non-convex functions might only reach a local rather than a global optimum.

3. **Bayesian optimization:** This method uses a probabilistic approach to predict the performance of various hyperparameters and updates the model as more results are observed (Snoek et al., 2012). The key idea is to balance exploration—testing hyperparameters where the model's predictions are uncertain—and exploitation—focusing on hyperparameters that are predicted to yield the best performance. In Bayesian optimization, a surrogate model, often a Gaussian process, is used to model the unknown function linking hyperparameters to an objective function, such as model accuracy. Gaussian processes are favored for their ability to provide a smooth estimate and naturally incorporate prediction uncertainty. An acquisition function, such as expected improvement, probability of improvement, or upper confidence bound, is then used to select the next set of hyperparameters to be evaluated, optimizing the use of the surrogate model (Frazier, 2018). Compared to methods like grid search, Bayesian optimization can identify more effective hyperparameters with significantly fewer evaluations (Injadat et al., 2018).

## Other Considerations

Considering that our goal is to learn and imitate the original data as precisely as possible—rather than capturing the underlying parameters of the overall population represented by the original data—an increase in model complexity and training duration might be beneficial. This approach allows us to capture and learn complex patterns and details within the data, thereby producing synthetic data that more closely resembles the original data.

However, it is essential to ensure that no specific samples from the original data, nor identifiable characteristics of particular individuals, are leaked. For example, a company has generated a synthetic dataset of employee information to research employee turnover prediction. The entries in this synthetic dataset include details such as years of service,

department, age, and salary range. Although the data are "synthetic" and do not exactly match actual data, some entries may closely resemble the characteristics of actual employees. For instance, if the synthetic dataset includes an entry showing an employee with 60 years of service, age 80, working in the IT department, and within a specific salary range, and if the actual dataset contains only one employee matching these conditions (the outlier), it could lead to speculation that this synthetic data entry was based on that particular employee's data.

To prevent generating samples that are identical or overly similar to the original data, we need to adequately randomize or obfuscate unique patterns within the original data. Additionally, methods such as data deduplication (e.g., Xia et al. 2016) should be used to reject any duplicate or excessively similar data.

## Declarations

**Competing Interests** The authors declare no competing interests.

## References

Abreu, S. (2019). Automated architecture design for deep neural networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1908.10714

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS ONE, 6*(9), e24357. https://doi.org/10.1371/journal.pone.0024357

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement? Does it matter? *Personnel Psychology, 51*(3), 577–598. https://doi.org/10.1111/j.1744-6570.1998.tb00252.x

Banks, G. C., Field, J. G., Oswald, F. L., O'Boyle, E. H., Landis, R. S., Rupp, D. E., & Rogelberg, S. G. (2019). Answers to 18 questions about open science practices. *Journal of Business and Psychology, 34*(3), 257–270. https://doi.org/10.1007/s10869-018-9547-8

Bemis, S. E. (1968). Occupational validity of the general aptitude test battery. *Journal of Applied Psychology, 52*(3), 240–244. https://doi.org/10.1037/h0025733

Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural Computation, 12*(8), 1889–1900. https://doi.org/10.1162/089976600300015187

Braddy, P. W., Gooty, J., Fleenor, J. W., & Yammarino, F. J. (2014). Leader behaviors and career derailment potential: A multi-analytic method examination of rating source and self–other agreement. *The Leadership Quarterly, 25*(2), 373–390. https://doi.org/10.1016/j.leaqua.2013.10.001

Caron, G., & Srivastava, S. (2022). Identifying and manipulating the personality traits of language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.10276

Carpenter, N. C., Son, J., Harris, T. B., Alexander, A. L., & Horner, M. T. (2016). Don't forget the items: Item-level meta-analytic and substantive validity techniques for reexamining scale validation. *Organizational Research Methods, 19*(4), 616–650. https://doi.org/10.1177/1094428116639132

Castille, C. M., Köhler, T., & O'Boyle, E. H. (2022). A brighter vision of the potential of open science for benefiting practice: A

ManyOrgs proposal. *Industrial and Organizational Psychology, 15*(4), 546–550. https://doi.org/10.1017/iop.2022.70

Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.20094

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., & Xiao, Y. (2024). From persona to personalization: A survey on role-playing language agents. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.18231

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLoS ONE, 15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Couture, J. L., Blake, R. E., McDonald, G., & Ward, C. L. (2018). A funder-imposed data publication requirement seldom inspired data sharing. *PLoS ONE, 13*(7), e0199789. https://doi.org/10.1371/journal.pone.0199789

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81–100. https://doi.org/10.1037/a0015914

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2014). Certifying and removing disparate impact. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1412.3756

Fleenor, J. W., Smither, J. W., Atwater, L. E., Braddy, P. W., & Sturm, R. E. (2010). Self–other rating agreement in leadership: A review. *The Leadership Quarterly, 21*(6), 1005–1034. https://doi.org/10.1016/j.leaqua.2010.10.006

Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. *Journal of Big Data, 10*(1). https://doi.org/10.1186/s40537-023-00792-7

Frazier, P., I. (2018). A tutorial on Bayesian optimization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1807.02811

Gabriel, A. S., & Wessel, J. L. (2013). A step too far? Why publishing raw datasets may hinder data collection. *Industrial and Organizational Psychology, 6*(3), 287–290. https://doi.org/10.1111/iops.12051

Gerpott, F. H., Lehmann-Willenbrock, N., Voelpel, S. C., & Van Vugt, M. (2019). It's not just what is said, but when it's said: A temporal account of verbal behaviors and emergent leadership in self-managed teams. *Academy of Management Journal, 62*(3), 717–738. https://doi.org/10.5465/amj.2017.0149

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. https://doi.org/10.1145/3422622

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1406.2661

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org/

Hamilton, D. G., Hong, K., Fraser, H., Rowhani-Farid, A., Fidler, F., & Page, M. J. (2023). Prevalence and predictors of data and code sharing in the medical and health sciences: Systematic review with meta-analysis of individual participant data. *BMJ*, e075767. https://doi.org/10.1136/bmj-2023-075767

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science : An

observational study. *Royal Society Open Science*, 8(1). https://doi.org/10.1098/rsos.201494

Huang, J., Wang, W., Li, E. J., Lam, M. H., Ren, S., Yuan, Y., Jiao, W., Tu, Z., & Lyu, M. R. (2023). Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench. *arXiv* (Cornell University). https://doi.org/10.48550/arxiv.2310.01386

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. Springer.

Injadat, M., Salo, F., Nassif, A. B., Essex, A., & Shami, A. (2018). Bayesian optimization with machine learning algorithms towards anomaly detection. *IEEE Global Communications Conference (GLOBECOM)*. https://doi.org/10.1109/glocom.2018.8647714

Jiang, H., Zhang, X., Cao, X., Kabbara, J., & Roy, D. (2023). PersonaLLM: Investigating the ability of large language models to express big five personality traits. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.02547

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic data - What, why and how? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.03257

Kamthe, S., Assefa, S., & Deisenroth, M. P. (2021). Copula flows for synthetic data generation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2101.00598

Kato, A. E., & Scherbaum, C. A. (2023). Exploring the relationship between cognitive ability tilt and job performance. *Journal of Intelligence*, 11(3). https://doi.org/10.3390/jintelligence11030044

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy, J. H., Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J., Lowery, S., Kitching, M., Nimbkar, A., Boyce, A., Sun, T., Guo, F., … Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*, 76(4), 1061–1123. https://doi.org/10.1111/peps.12608

Kreamer, L. M., Albritton, B. H., Tonidandel, S., & Rogelberg, S. G. (2023). The use and misuse of organizational research methods 'best practice' articles. *Organizational Research Methods*, 26(3), 387–408. https://doi.org/10.1177/10944281211060706

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. *Springer eBooks*. https://doi.org/10.1007/978-1-4614-6849-3

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7). https://doi.org/10.1093/pnasnexus/pgae233

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, 41(11), 1183–1192. https://doi.org/10.1037/0003-066x.41.11.1183

Leavitt, K. (2013). Publication bias might make us untrustworthy, but the solutions may be worse. *Industrial and Organizational Psychology*, 6(3), 290–295. https://doi.org/10.1111/iops.12052

Lee, A., & Carpenter, N. C. (2018). Seeing eye to eye: A meta-analysis of self-other agreement of leadership. *The Leadership Quarterly*, 29(2), 253–275. https://doi.org/10.1016/j.leaqua.2017.06.002

Leslie, J., & Braddy, P. W. (2015). *Benchmarks for managers technical manual*. Center for Creative Leadership.

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.07849

Meriac, J. P., Hoffman, B. J., & Woehr, D. J. (2014). A conceptual and empirical review of the structure of assessment center dimensions. *Journal of Management*, 40(5), 1269–1296. https://doi.org/10.1177/0149206314522299

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

Nowok, B., Raab, G. M., & Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R1. *Statistical Journal of the IAOS*, 33(3), 785–796. https://doi.org/10.3233/sji-150153

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. https://doi.org/10.1109/dsaa.2016.49

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916231214460

Pew Research Center. (2019). *Trust and mistrust in Americans' views of scientific experts*. https://www.pewresearch.org/science/2019/08/02/trust-and-mistrust-in-americans-views-of-scientific-experts/

Porter, C. C. (2008). De-identified data and third party data mining: The risk of re-identification of personal information. *Shidler JL Com. & Tech.*, 5.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raghunathan, T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, 8(1), 129–140. https://doi.org/10.1146/annurev-statistics-040720-031848

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (2.4.6.24) Retrieved from https://doi.org/10.32614/cran.package.psych

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, 9(2), 461–468.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation. *Organizational Research Methods*, 19(1), 80–110. https://doi.org/10.1177/1094428115598239

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1206.2944

Tang, R., Han, X., Jiang, X., & Hu, X. (2023). Does synthetic data generation of LLMs help clinical text mining? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.04360

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8(1). https://doi.org/10.1038/s41597-021-00981-0

Towse, J. N., Ellis, D. A., & Towse, A. S. (2020). Opening Pandora's Box: Peeking inside psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*, 53(4), 1455–1468. https://doi.org/10.3758/s13428-020-01486-1

U.S. Department of Labor. (1970). *Manual for the uses General Aptitude Test Battery. Section III: Development*. Washington, DC: U.S. Government Printing Office.

United Kingdom Statistics Authority. (2022). *Ethical considerations relating to the creation and use of synthetic data*. https://uksastatisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and%20use-of-synthetic-data/

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, 1(1). https://doi.org/10.1525/collabra.13

Venables, W., & Ripley, B. (2002). *Modern applied statistics with S*. Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the General Aptitude Test Battery. *Journal of Applied Psychology, 78*(6), 981–987. https://doi.org/10.1037/0021-9010.78.6.981

Wang, P., Xiao, Z., Chen, H., & Oswald, F. L. (2024a). Will the real Linda please stand up. . .To large language models? Examining the representativeness heuristic in LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.01461

Wang, P., Zou, H., Yan, Z., Guo, F., Sun, T., Xiao, Z., & Zhang, B. (2024b). Not yet: Large language models cannot replace human respondents for psychometric research. *OSF*. https://doi.org/10.31219/osf.io/rwy9b

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*(7), 726–728. https://doi.org/10.1037/0003-066x.61.7.726

Xia, W., Jiang, H., Feng, D., Douglis, F., Shilane, P., Hua, Y., Fu, M., Zhang, Y., & Zhou, Y. (2016). A comprehensive study of the past, present, and future of data deduplication. *Proceedings of the IEEE, 104*(9), 1681–1710. https://doi.org/10.1109/jproc.2016.2571298

Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1811.11264

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.1907.00503

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zhang, L., Wu, Y., & Wu, X. (2016). A causal framework for discovering and removing direct and indirect discrimination. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1611.07509