SOCIETY REGISTER

Original Article

The hidden problem in Big Data: even infinite information does not guarantee consistent measurement

SOCIETY REGISTER
2024 / 8(4): 7–30
ISSN: 2544–5502
DOI: 10.14746/sr.2024.8.4.01

OPEN ACCESS

Dino Carpentras[1] & Philip Warncke[2]

[1] ETH Zurich, 8092, Stampfenbachstrasse 48, Zurich, Switzerland. ORCID: 0000-0001-8471-2352, Email: dino.carpentras@gmail.com

[2] Freie Universität Berlin, Kaiserswerther Str. 16-18, 030 8381, Berlin, Germany. ORCID: 0000-0001-7939-1267, Email: p.warncke@fu-berlin.de

ABSTRACT: The social sciences heavily depend on the measurement of abstract constructs for quantifying effects, identifying associations between variables, and testing hypotheses. In data science, constructs are also often used for forecasting, and thanks to the recent big data revolution, they promise to enhance their accuracy by leveraging the constantly increasing stream of digital information around us. However, the possibility of optimizing various social indicators implicitly hinges on our ability to reliably reduce complex and abstract constructs (such as life satisfaction or social trust) into numeric measures. While many scientists are aware of the issue of measurement error, there is widespread, implicit hope that access to more data will eventually render this issue irrelevant. This paper delves into the nature of measurement error under quasi-ideal conditions. We show mathematically and by employing simulations that single measurements fail to converge even when we can access progressively more information. Then, by using real-world data from the Social Capital Benchmark Surveys, we demonstrate how adding new information increases the dimensionality of the measured construct quasi-indefinitely, further contributing to measurement divergence. We conclude by discussing implications and future research directions to solve this problem.

KEYWORDS: measurement, Big Data, measurement error, measurement crisis, simulation, Social Capital Benchmark Surveys (SCBS)

> *Absolute space, in its own nature, without regard to anything external, remains always similar and immovable. Relative space is some movable dimension or measure of the absolute space, [...] which is vulgarly taken for immovable space.*
>
> — Sir Isaac Newton, 1689

INTRODUCTION

The social sciences heavily rely on the measurement of abstract concepts, such as happiness, social trust, or political ideology, which cannot be observed directly but must be inferred from revealed data. Besides their immediate value to science, measurement models are also heavily used in applied technology, ranging from innocuous online personality tests to more concerning examples, such as China's citizen scores. Such technology is often complemented by new approaches from computational social science, such as AI (O'Leary, 2013; Gligor et al., 2021), digital twins (Li et al., 2022) and large-language models (LLMs), which all promise to turn big data into valuable knowledge (Yenduri et al., 2024). The rise of big-data science indeed belies a certain optimism for social engineering. In pursuit of the credo, "what can be measured can be improved" (Krummel, 2019), we might envision a society in which computational social science, equipped with big data, generates insights capable of enhancing almost all aspects of society.

However, we argue for caution. Data does not automatically translate to useful knowledge; a critical first step in analyzing behavioral and attitudinal data involves converting raw information into measures of theoretically meaningful concepts. Our ability to improve society thus depends on how well we can measure the desired outcomes. For example, as smiling is associated with happiness across many cultures (Szarota, 2011), it is not too far-fetched to imagine a slightly dystopian future in which facial recognition software attempting to measure citizens' happiness simply focuses on how often people are smiling. In this case, some individuals or groups may be incentivized to artificially boost their happiness scores by smiling more often, even if this may negatively affect their mood (Labroo et al., 2014).

Scholars have long been aware of fundamental issues associated with the measurement process and have proposed various strategies for addressing them (Spearman, 1904; Stevens, 1946; Rash, 1960; Bollen, 1980). Measurement error—arguably the most well-studied among these (Challen et al., 2019; Roselli et al., 2019; McNamara et al., 2022)—is often conceptualized in such a way that it progressively disappears as we gain access to more and more information. The fundamental problem of measurement is thus often portrayed as one of limited access to information. In the words of Bandalos (p. 4, 2018), the basic limit to measurement precision lies in the fact that "*we cannot ask every possible question or observe every instance of behavior*". But what if the big data revolution indeed

gets us closer to numerically tracking every instance of human behavior? Would infinite information strip away all inconsistencies in the measurement of our desired concepts? Using simulations and an applied example, we show that this is not the case. Throughout this article, we elaborate on more fundamental problems rooted within the measurement process that require conceptual solutions that data alone cannot provide.

In the next section, we delve more deeply into the main problems of measuring social science concepts. We will then investigate if measurement error truly disappears in the limit of infinite information. We approach this issue threefold: first mathematically, then by using simulated data, and finally with empirical data from the Social Capital Benchmark Surveys. Notice that our purpose here is not to repeat the same analysis at three different levels. Instead, we take advantage of our findings at each level to shed different light on the same problem.

THE PROBLEMS OF MEASUREMENT

THE THEORETICAL PROBLEM

In the physical world, measures are often based on measurement units. In order to have a practical meaning, measurement units require two qualities called **equivalence** and **concatenation** (Krantz et al., 1971, 1989, 1990). **Equivalence** allows us to determine if two objects possess the same quantity of the construct we intend to measure. In comparison, **concatenation** refers to the ability to combine two objects to form a new one. For example, we might select a rod and designate it as our unit of length. If we find another rod of equal length (equivalence) and combine both rods (concatenation), we create a new rod that is exactly twice as long as the original. Thus, any object being as long as this new rod can be measured as "two rods long". Similarly, anything 7 inches long is equivalent to seven concatenated rods, each 1 inch long.

Unfortunately, equivalence and concatenation are impossible operations for the concepts we tend to measure in the social world. Even if we decide that one person serves as our baseline unit for life satisfaction, no known logical operation allows us to unambiguously combine her life satisfaction with another person's life satisfaction to obtain an individual who is twice as satisfied. This means that key properties of physical measurements cannot be directly transplanted into the social world. We must rely on other approaches or invoke additional theories and assumptions to measure social concepts.

Scholars have proposed a number of theories to measure abstract concepts (DeVellis, 2006; Hambleton et al., 1991). However, the only one that partially solves the issue of absent measurement units is the Additive Conjoint Theory (ACT) (Luce, 1966; Krantz et al., 1971). However, due to its highly stringent data requirements, ACT is almost never used in practice. All other statistical approaches are based on modeling abstract concepts as numeric, latent variables or constructs, including Classical Test Theory (DeVellis, 2006), factor analysis (Spearman 1904; Bollen 1980), and Item Response Theory (Hambleton et

al., 1991; Kean & Reilly, 2014). This has prompted several scholars to criticize latent variable models, claiming that they perform *data reduction* instead of actual *measurement* (Uher, 2021; Thompson & Vacha-Haase, 2000). However, even if many researchers are aware of the theoretical limitations of latent variable models, there is a strong consensus that they are useful in practice (e.g., Kline, 2023, p. 414f.).

While this may seem contradictory, using incomplete or "problematic" models has been quite common across scientific disciplines. For example, mathematical tools like the Dirac delta function have long been used despite possessing logical inconsistencies in their initial formulation. It was only with the advent of distributions theory that the Dirac function was made mathematically consistent (Halperin & Schwartz, 1952). Moreover, although Heisenberg's uncertainty principle poses a hard, upper limit to the maximum precision achievable with physical measurements (Heisenberg, 1927), physics and engineering kept advancing our understanding of nano-world phenomena (Sanchez & Sobolev, 2010). Likewise, many cultures were able to construct elaborate buildings before inventing writing systems (Webster, 1996), which implies that they most likely lacked sophisticated theories of mechanics.

In sum, both the nature and evolution of measurement practice in the natural sciences suggest that the social sciences can incrementally improve measurement practice even if logical inconsistencies and theoretical problems—such as a lack of equivalence and concatenation—persist. In the next section, we will explore if the problem of measurement validity is simply confined to a lack of theoretical closure or if there are more fundamental, practical limitations.

PRACTICAL PROBLEMS

Theoretically speaking, it is always possible to map objects to numbers. For example, we could assign a random number to each revealed preference in a social survey. However, such a process would certainly not result in valid measurement. First, we need to establish that assignments of increasing numeric values correspond to some form of growing intensity, quantity, or abundance of the underlying concept—a quality sometimes called numeric correspondence. Second, we need to ensure that our numeric mapping indeed captures the target concept—a property referred to as conceptual correspondence. Researchers have proposed the concept of measurement validity (Nunnally, 1978; Carmines & Zeller, 1979), which is achieved when a measurement process possesses both numeric and conceptual correspondence. For instance, a valid life satisfaction scale assesses life satisfaction and not something else, like gross annual income or the severity of back pain. Likewise, someone with a "3" on this scale possesses lower life satisfaction than someone with a "10".

However, measurement validity does not ensure that the construct under scrutiny can itself be reduced to a single value. To better understand this problem, let's pose the fol-

lowing slightly absurd question: "*What is the language spoken by Europeans?*" Of course, one should recognize that the question is ill-posed as it suggests that only one language is spoken in Europe. However, for the sake of exposition, let's imagine a researcher who tries to answer this question anyhow. One research strategy might involve visiting the geographical midpoint of Europe. Since this point lies in Gadheim, Germany, the scientists will most likely converge on the conclusion that German is the language spoken by Europeans. As a validity check, further analysis might confirm that the point chosen for the measurement is indeed in Europe; thus, the researcher was ostensibly measuring what they initially intended to measure.

While the above example might appear absurd, it exhibits significant similarities with measurement processes. On a fundamental level, measurement can be thought of as answering the question, "What is the number that better represents the construct within examined data?" Validity checks ensure a connection between the numeric assignment and the concept, but there is no real fail-safe if the question implied by the measurement process does not make sense to begin with.

Applied researchers may interject that this verification process can be carried out using metrics like Cronbach's alpha (Cronbach, 1951) or statistics assessing the underlying dimensionality of a given input data set (Horn, 1964). However, it is important to stress that such metrics are properties inherent to the data and not the concept itself. We can perform statistical tests to assess whether a given dataset can be summarized reasonably well by a single (or finite number of) numeric variable(s). However, we have no proof that such a variable truly and uniquely represents the underlying concept.

One exciting aspect of our previous example is that if another scientist tries to answer the same question (i.e., *what is the language spoken by Europeans?*), they may choose a different measurement procedure. Instead of the geographic center of Europe, the researcher may instead focus on the most widely spoken native language on the European continent, which will lead them to obtain a different—yet reproducible—answer (i.e., Russian). In this way, we can end up with the paradoxical situation of having two valid answers to the same question. Even more concerning is that it is possible to multiply such an approach by continuously re-operationalizing and testing for validity to ensure that the target construct can truly be reduced to single numeric measures.

In recent years, scientists have started employing similar strategies through the lens of "researcher-degrees-of-freedom". Such studies attempt to answer the same research question using different numeric operationalizations, often finding starkly contrasting results (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018; Landy et al., 2020; Breznau et al., 2022; Carpentras, 2024; Warncke et al., 2024). Similarly, researchers have tested the consequences of order-preserving transformations within measurement models. This work has shown that it is possible to invert many published results (Schröder & Yitzhaki, 2017) and introduce uncertainty as large as the entire range of the outcome space (Car-

pentras & Quayle, 2023). However, each of these studies has been conducted in the case of limited data, meaning there is no hard proof that these effects are not simply due to measurement error.

In the following sections, we will further explore the possibility of measurement convergence in the "ideal" case of increasingly available data. For our analysis, we thus formulate the following general hypothesis:

> (1) *In the case of infinite data, measurements of the same concept on the same population are independent of data.*

To evoke a physical analogy, the height of an object does not change when we change our measurement instrument. Since hypothesis 1 cannot be tested using simulations or data (i.e., we cannot collect or simulate infinite data), we also produce another, more testable hypothesis which nevertheless captures the same spirit of our initial formulation:

> (2) *As the available data on the same population and same construct increases, all measurements of latent constructs should converge.*

Notice that this definition is consistent with how the concept of infinity is defined in mathematics. Indeed, the fact that goes to 0 when goes to infinity is not conceptualized as the fact that can ever be equal to infinity. Instead, this is conceptualized as the fact that for bigger and bigger values of (i.e., as "approaches" infinity), the value gets closer and closer to 0 (Courant et al., 1965).

LATENT VARIABLE METHODS

Researchers have proposed a number of methods and associated theories that attempt to extract latent variables from data, including classical test theory, factor analysis, and item response theory (DeVellis, 2006; Bollen, 1980; Hambleton et al., 1991). However, the equations proposed by these theories are not derived from behavioral patterns or studies of the human brain. Instead, they have been simply suggested as theories, making them valuable tools for summarizing data while providing no built-in guarantee for measurement validity.

Classical test theory models the measurement process according to the following formula (DeVellis, 2006):

Equation 1

$$x = \frac{1}{N}\sum_{i}^{N} r_i = \frac{1}{N}\sum_{i}^{N}(t + \varepsilon_i)$$

Where x is the result of the measurement process, r_i is the response level to item i, N is the number of collected items, t is the accurate score (i.e., the real value of the construct),

and ε_i is the randomly distributed error with a mean equal to 0. Under these assumptions, as *N* increases, the error term averages out. Hence, measurements will coincide with the true value *t*. Note that infinite information (i.e., items) must lead to perfect measurement within this theoretical framework. However, this is only true if people behave precisely as the underlying theory hypothesizes. Also, notice that the term "error" is misleading, as it suggests that when people report their responses, they report the true value plus some degree of measurement error. According to this model, someone's opinion on immigration, for instance, should be considered their value on the left-right ideological spectrum plus some "error".

To better explore convergence, we can use the above-introduced re-operationalization test. Specifically, we can think of a hypothetical scientist selecting items from the same pool in two separate ways. One fraction of 1-f of items is chosen uniformly, so that the mean value of the error is equal to 0. However, a different fraction, f, is more correlated, so the mean error is not zero.

As a more practical example, we may consider a researcher attempting to measure the left-right political spectrum. A fraction 1-f of items uniformly spans over multiple topics, such as gun control, immigration, women's rights, etc. Instead, a fraction f of them focus on a specific topic, such as gay rights. While we can still assume that the average error is 0 for the uniformly chosen topics, it does not make sense to assume the same to be true for the items related to the same topic. Indeed, if we give a participant 10 items on gay rights, we do not expect them to produce the same result as if we were proposing 10 items on other political topics. This is because there is no reason to believe that people's opinions on gay rights are exactly equal to their position in the left-right spectrum. Thus, if we keep referring to *t* as the true value and *v* as the average responses to the selected topic, we modify equation 1 into:

Equation 2

$$x = \frac{1}{N}\left(\sum_{i}^{N(1-f)} r_i + \sum_{j}^{Nf} r_j\right) \sim (1-f)t + fv$$

In the final term, we applied the case of infinite N.

Notice that this approach belies that the outcome of the measurement process depends on the data selection process—even in the case of infinite information. To make this point even more precise, suppose that two researchers select items among different fractions f. For instance, researcher 1 may prefer items related to women's rights, while researcher 2 over-selects gay rights items. From equation 2, we can infer that in the limit of infinite N, researcher 1 would obtain ~ (1- *f*)t + fv_2. Thus, the difference between the two measurements would be:

Equation 3

$$x_1 - x_2 = f(v_1 - v_2)$$

The only way to make this difference equal zero is either by having $f = 0$ or $v_1 = v_2$. This either supposes that all topics are perfectly equivalent to one another (i.e., asking about women's rights is equivalent to asking about gay rights) or that the two researchers have chosen exactly the same method for fielding and selecting the items they did. Since these assumptions are rarely defendable in social science research, we cannot obtain the same results, even if all researchers relied on the same theoretical framework, used the same population, and had access to infinite items. This conclusion stands in stark contrast to hypothesis 1.

Some readers may object that different choices of items might introduce undue bias. Indeed, the two researchers are not using the same "correct" set of items, but they are indulging too much on certain topics and are, therefore, biasing the measurement process. However, this begs the question as to what the correct choice of items should be to begin with.

If we could unambiguously point to a single, "correct" set of items, all scientists would always (re-)produce the same result. In this case, would always be equal to which would indeed obviate the need for a consistent measurement framework. However, if researchers cannot agree on a single, optimal set of items, we need to accept that different scientists will make different selections in the real world. Unfortunately, even if they keep collecting more and more data, they will thus never achieve measurement convergence. In the next section, we will further explore the consequences of this specific case using simulations.

SIMULATIONS

Simulations allow us to see better how the above-outlined measurement theories behave under increasing volumes of data. We avoid imposing any particular measurement model or theory to generate our simulation data. We opted for this strategy because (1) different measurement theories do not agree with each other because they rely on different assumptions about the data-generating process, and (2) if we created data based on a specific measurement model, we would trivially obtain consistent results because they are based on the same theoretical framework.

Using an 'a-theoretical' method, we simply generated correlated response patterns. In particular, we simulated 1,000 respondents providing answers to 50,000 hypothetical survey questions, each featuring 5 response levels. These response patterns were generated from an initial set of 1,000 random numbers drawn from the uniform interval

between 1 and 5. We refer to this array as V_1. Then we generated every V_i by adding noise in the form of a uniform random variable of amplitude 0.0018 to V_{i-1} while bounding the results in the interval [1, 5]. We then repeated this process 49,999 times. Finally, each array of items was obtained by simply rounding the corresponding V_i.

The average correlation between our simulated items is 0.88, while we estimated the lowest correlation between any item pair at 0.73; our items are thus statistically very similar. Using factor analysis, we consistently identify only a single eigenvalue above magnitude 1 (see Figure 1). Importantly, even though these data were not generated based on any particular measurement theory or model stipulating the existence of a single latent construct, the scree plot in Figure 1 strongly suggests that a single latent factor might have produced the data.

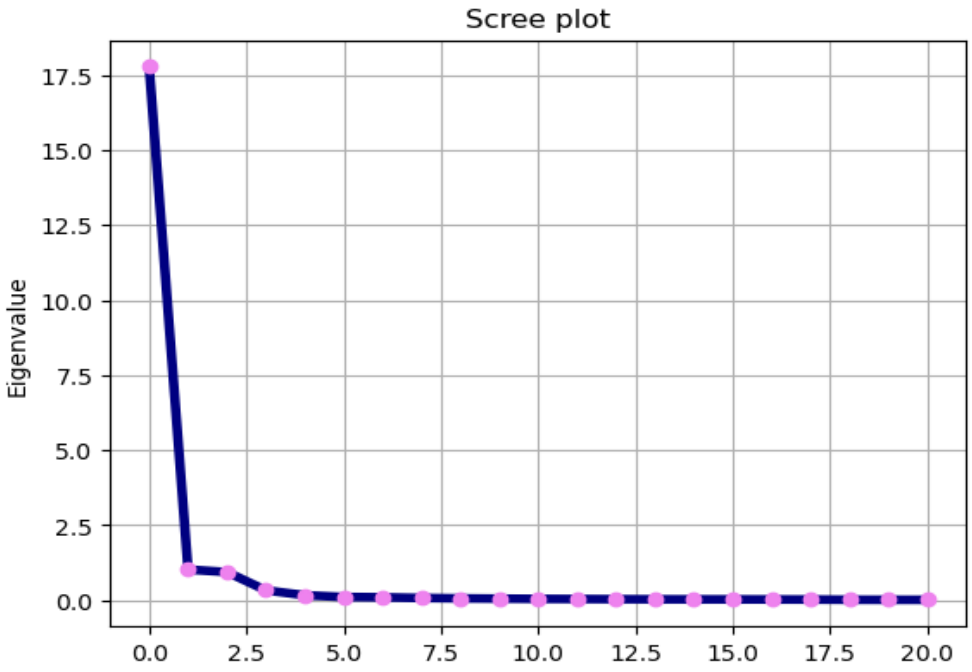

Figure 1. Scree plot showing uni-dimensionality of the simulated data

What would happen if two separate researchers were to use different item subsets from this data pool? Suppose our hypotheses were true as more and more data is collected. In that case, measurements of latent constructs should increasingly become independent from any particular section of items. In practice, this would mean that even if two scientists do not select the exact same item set, their measurements should converge as data availability increases. Conversely, suppose the measurement process is limited to sheer dimensionality reduction. In that case, if numeric assignments are not consistently capturing a single underlying concept, different item selections would prevent measurement convergence.

To model variability in the process of item selection, we specified two linear probability

distributions. In the first, item 1 has the maximum probability of being selected, while item 50,000 possesses the lowest. These values were simply reversed in the second probability distribution,

Our simulation of the measurement process proceeds as follows:

(i) Scientists 1 and 2 select N random items from the item pool using the two opposite selection distributions.

(ii) Both scientists extract a single latent variable using factor analysis and classical test theory from their respective data selections.

(iii) Each scientist calculates the score of each participant using both methods.

(iv) Each scientist ranks the respondents based on scores obtained by the latent variable models.

We repeated this experiment by varying N (the number of items) and repeated the same configuration 10 times for each N. This allows us to calculate both the mean and the standard deviation for each value of N. To establish a baseline benchmark comparison, we repeated the same analysis for the case where both scientists used the same sampling distribution.

In Figures 2 and 3, we report a number of summary statistics, relying on the factor analysis implementation in FactorAnalyzer from Python's sklearn (Pedregosa et al., 2011) using "varimax" as rotation. The four sub-figures show:

(a) the correlation between the latent factors obtained by the two scientists

(b) the number of people who share the same score ranking for the two scientists

(c) the average rank difference

(d) the fraction of people scoring in the top 5% of both rankings

The average rank difference is defined as follows:

$$m_r = mean\left(\frac{r_{1,j} - r_{2,j}}{R}\right)$$

Where $r_{i,j}$ is the ranking of participant j according to scientist i, and R is the maximum possible rank (i.e., the total number of respondents).

Figure 2 summarizes the scenario where both scientists use the same item distribution. Here, as the number of measurements increases, the two scientists are more likely to select the same items, and the two factors produce more and more similar results across all metrics. However, things look very different if the two scientists do not rely on the same

sampling distribution. While all measurements show an initial improvement in Figure 3, this improvement rapidly saturates. This process illustrates that the two scientists ultimately measure two different latent constructs.

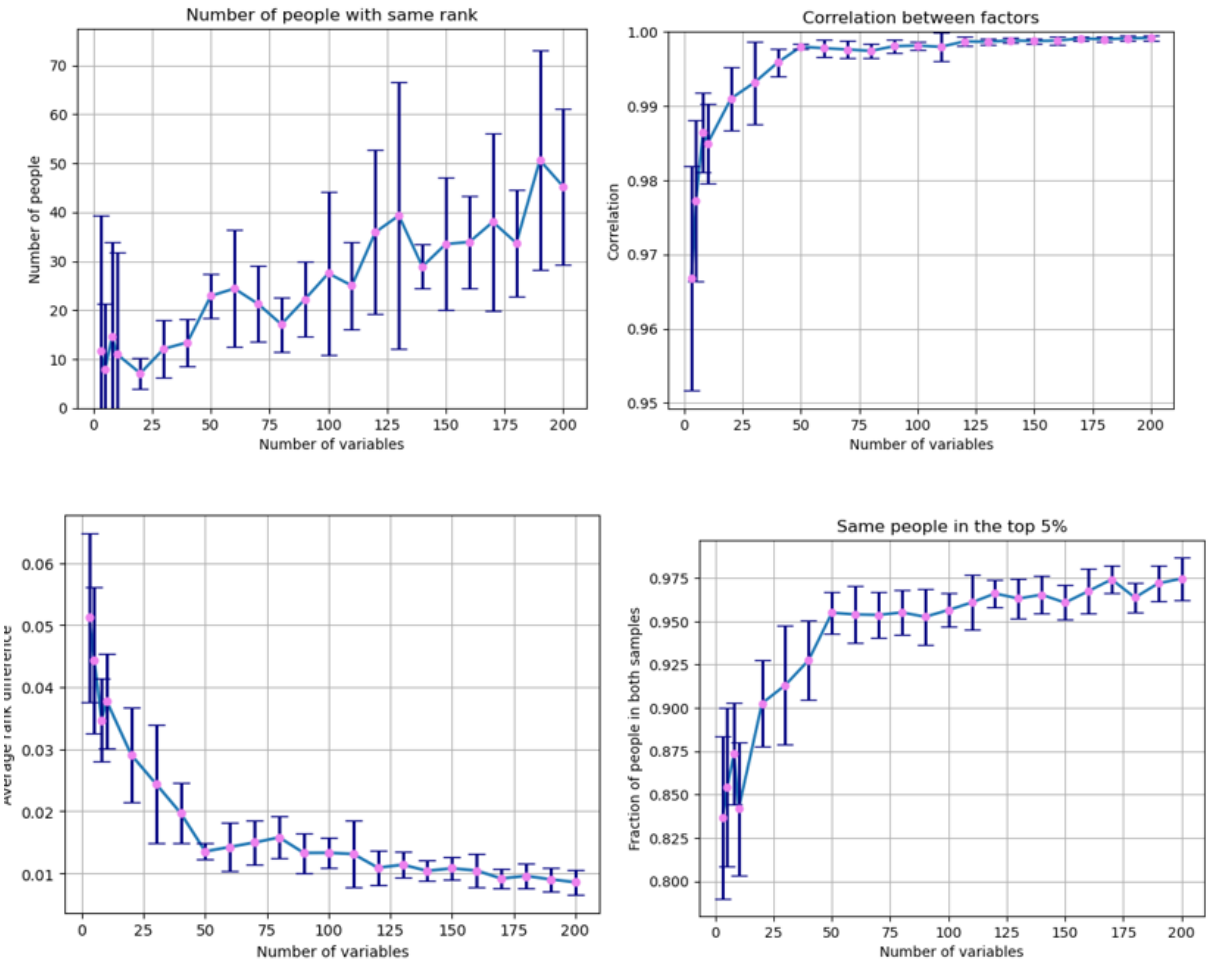

Figure 2. Same item sampling distributions.
Measurements converge as the number of selected items increases

Some readers may object that the above results may not classify as wrong measurements by social science standards. Indeed, even in the "worst case", the two scientists produce latent quantities which are correlated at up to r = 0.98. Metrics of this magnitude are almost never achieved in practice. But is this really a worst-case scenario?

For context, we were only able to achieve this level of convergence by assuming that the two researchers have (1) access to data from the exact same population, (2) were able to collect up to 200 items which (3) exhibit an average correlation of 0.88 and (4) analyzing the dataset with the exact same method. Despite all these advantages, if the two researchers were asked to select the people who scored in the top 5%, they would still disagree with one another about 15% of the time.

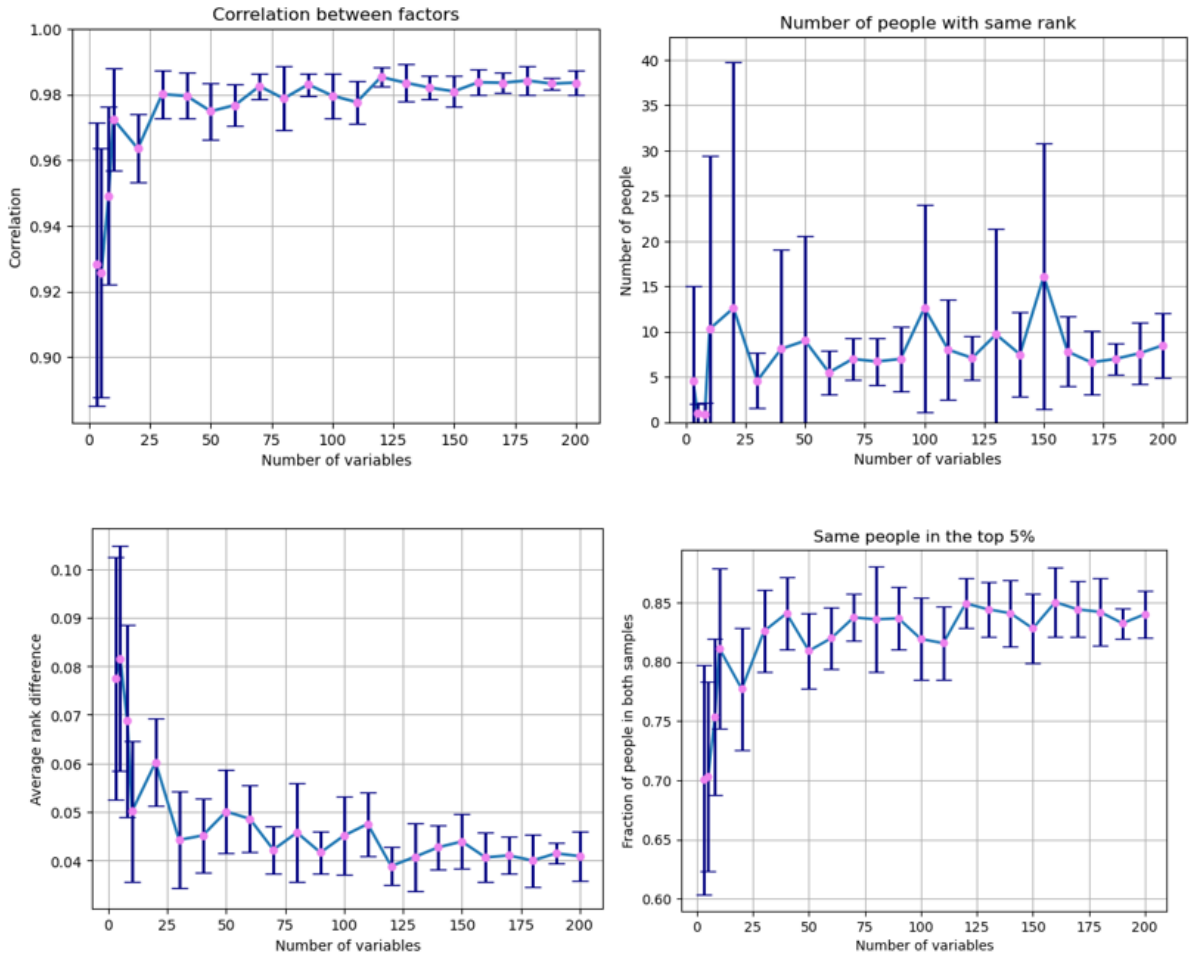

Figure 3. Different item sampling distributions.
Measurements ultimately fail to converge as the number of selected items increases.

However, the crucial point here does not relate to the absolute size of the remaining measurement error, but the lack of true measurement convergence. Indeed, our simulation shows that even in the case of quasi-boundless data availability, results obtained from these methods remain bounded to the particularities of each dataset. Strictly speaking, this violates hypothesis 2. In other words, the above measurement methods perform very well in terms of reducing the dimensionality inherent in a given data-set. Still, they do not provide true measurements of underlying constructs and are ultimately bound to arbitrarily selecting items.

EMPIRICAL APPLICATION: THE SOCIAL CAPITAL BENCHMARK SURVEYS

Thus far, we have analyzed measurement quality under the assumption of uni-dimensionality. However, researchers often measure multi-dimensional concepts. For example, personality is commonly divided into five main dimensions, also known as "the big 5" (Costa & McCrae, 1992). In this section, we rely on a real-world example to explore if multi-dimensionality might solve the measurement problem by guaranteeing conver-

gence among different conceptual sub-dimensions.

We focus on the Social Capital Benchmark Survey (SCBS)—an exceptionally large and widely used data source featuring individual-level survey responses to more than 60 items about volunteer engagement, local life, and sense of community. The dataset consists of 38 individual surveys of select US communities and an additional national probability sample.

One of the principal aims of the SCBC was to improve the measurement of *social capital*. The SCBC defines this concept as "[...] the value inherent in friendship networks and other associations which individuals and groups can draw upon to achieve private or collective objectives" (p. 3.). Like its economic counterpart, social capital has long been assumed to cause positive externalities among communities that possess it, such as higher economic growth, better schooling, higher voter turnout, and improved life satisfaction (c.f. Putnam, 1995; Partha & Serageldin, 2000). Not surprisingly, recent urban planning research has focused considerable attention on how smart cities can enhance social capital among residents (e.g., Caragliu et al., 2011; Kourtit & Nijkamp, 2012; Nakano & Washizu, 2021).

However, many empirical hypotheses about the drivers and consequences of social capital remain controversial among academic researchers; a recent meta-analysis about a purported negative relationship between social trust (a sub-dimension of social capital) and ethnolinguistic/ethnocultural diversity, for instance, finds large effect size heterogeneity across 81 studies spanning more than a dozen countries (Dinesen et al., 2020). We think this type of academic disagreement is at least partially rooted in issues of measurement inconsistency.

If a latent construct cannot be measured consistently, scholarly communities—at best—face obstacles in accumulating knowledge. At worst, they can get stuck in quasi-ideological debates about the supposed superiority of one set of measurement procedures over another. This is a particularly pernicious issue if different survey itineraries systematically bias research findings in a direction that one group of researchers finds desirable while another does not.

SAMPLING AND MODELING PROCEDURE

We identified 66 survey items in the SCBC used in previous research to measure social capital. The overall item set includes survey questions directed at people's sense of community, their levels of socio-political trust, the extent and diversity of social networks, pro-social attitudes, and self-reported volunteer activities[1]. Similar to our analysis above,

[1] Based on the major themes these items relate to (trust, volunteering, communal sense, and pro-sociality), researchers might reasonably expect at most four or five positively correlated of social capital which captures the data structure much more faithfully compared with a single underlying dimension. Howev-

we conducted a large number of statistical simulations by randomly selecting different pools of items from the social capital item population.

We initiated each of our simulations by drawing a random item sample of pre-specified size k. Next, we estimated the optimal number of latent dimensions needed to adequately reproduce the observed correlation patterns within this sample (Reckhase, 1990). Here, we relied on a state-of-the-art, machine-learning optimized procedure called Exploratory Graph Analysis (Golino & Epskamp et.al., 2017; Golino et.al., 2021). This method first finds the optimally sparse representation of the original correlation matrix using a graphical Gaussian Least-Absolute-Shrinkage (G-LASSO) algorithm. In the second step, the resulting network of inter-item correlations is passed to a community detection algorithm, which seeks to restore the optimal number of latent item clusters within the data sample. The thus obtained item communities are equivalent to latent factors in classic latent variable models; community membership can be interpreted as item loading patterns in confirmatory factor analysis (CFA, ibid)[2].

Equipped with a target number of dimensions and loading patterns, we subsequently fit CFA models using weighted least squares based on pair-wise complete polychoric correlation matrices to account for the ordered-categorical nature of and patterns of missingness in the data. We test all models for convergence and exclude those that fail to meet a relatively conservative convergence threshold (about 5% of the total sample)[3].

We repeat this process 200 times for each item bucket size, k, and progress the simulations by gradually increasing k until we reach a final bucket size of 64 items. This results in roughly 12,000 latent variable models containing nearly 70,000 latent social capital factors.

INFINITE DIMENSIONS, WEAK CONSISTENCY, LARGE EFFECT HETEROGENEITY

We visualize the results of our simulation in Figure 5. Panel A plots every model as a function of k (x-axis) and the optimal number of underlying latent dimensions. Our results reveal a stark linear dependency between both features. As we increase the number of sampled items, we are more likely to find additional latent dimensions of social capital ($r = 0.79$, $p < 0.01$). Panel A further shows that this relationship seems to grow without bounds. Panel A also reveals that the variance of latent dimensionality grows

er, as we will show in the results section, latent dimensionality grows without bound as we increase our item sample.

[2] Golino and Epskamp et.al. (2020) show in an extensive simulation study that EGA outperforms all conventional dimensionality detection techniques for ordered categorical data – matching our present data structure. However, a robustness-check using parallel analysis (Horn, 1964), a much simpler, re-sampling based strategy finds the same pattern of results.

[3] We treat a model as converged if the root mean square of the difference between the old and new parameter values are smaller than 1^-5. See *lavaan* package manual_section "lavOptions" for more details.

with item sample size, implying that uncertainty about the underlying factor structure increases as we ascertain more information about it ($r = 0.98$, $p < 0.01$). In this case, more information does not result in measurement convergence. If anything, more information increases the uncertainty about what we are, in fact, measuring.

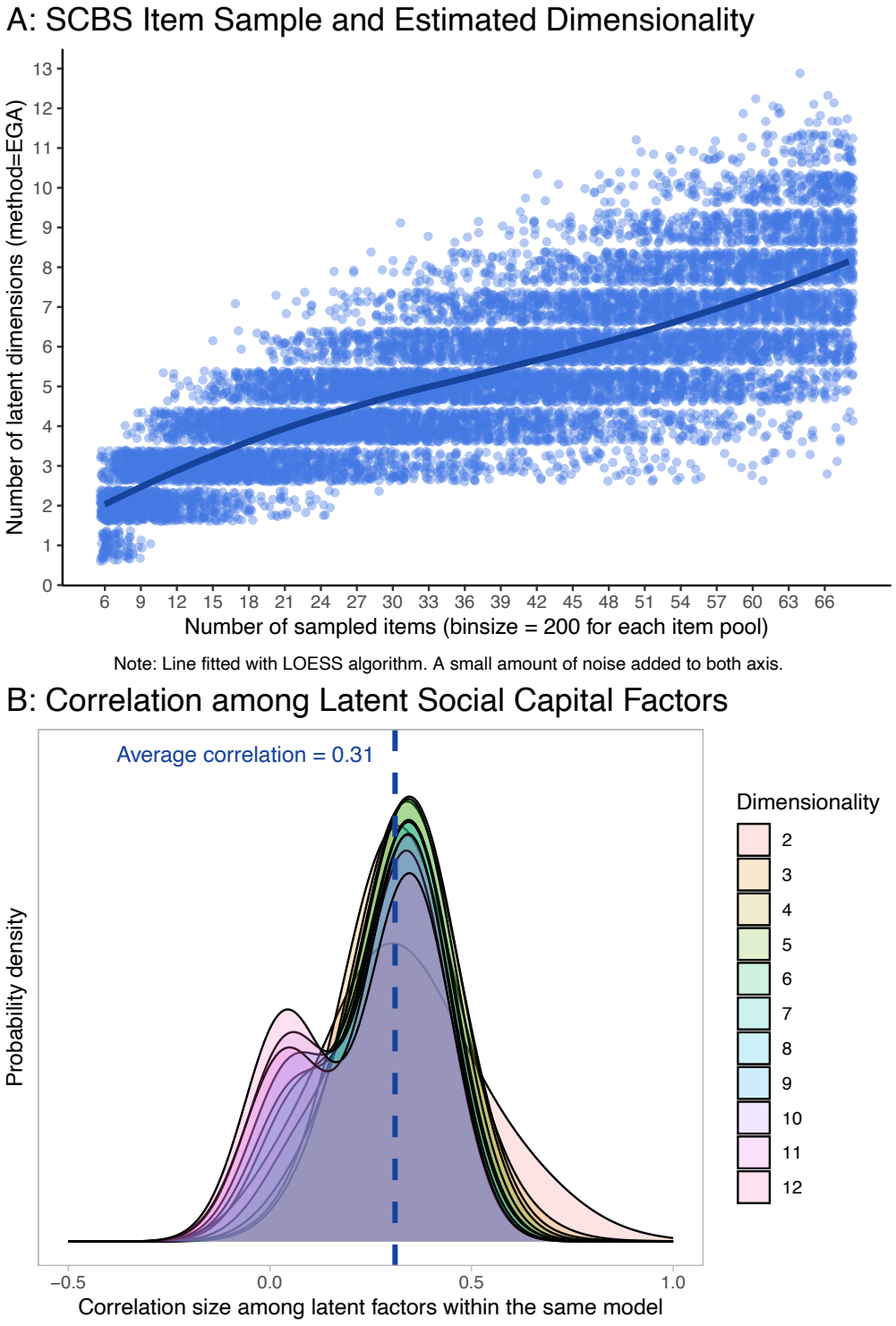

Figure 4. Latent dimensionality of social capital (SCBS 2000)

It is worth noting here that we uncovered this association only thanks to a continued process of re-operationalization. Without this extra step, researchers would instead likely

continue with their analysis and produce models with a specific, finite number of latent dimensions. Furthermore, other researchers repeating the same analysis on the same dataset would similarly, but mistakenly, confirm that the dataset contains the exact same number of latent social capital dimensions.

Some readers might question if the higher-dimensional solutions in Panel A are, in fact so closely related that they offer little to no additional insight. If we know one social capital factor that is part of, say, a six-dimensional solution in which all factors highly correlate with one another, there is little practical and theoretical value to interpreting, retaining, and using all six. In this case, higher dimensions could be little more than a nuisance artifact of the simulation procedure itself. Conversely, others may wonder if the population of survey items used here was well-suited to capture a single underlying construct to begin with. If the SCBC data are only weakly and arbitrarily correlated with one another, we should not take any solution, especially multi-dimensional ones, seriously. In this scenario, we would expect the mean factor correlation to be centered at zero and expect to find as many positively as negatively correlated factor pairs.

In short, two fundamental objections to our results in Panel A should manifest in opposite ways. For example, our conclusions would be questionable if the latent factor solutions were strongly positively correlated (i.e., all solutions essentially capture the same information) or not correlated (i.e., all solutions capture highly idiosyncratic information). Interestingly, Panel B demonstrates that neither is the case; based on the probability densities of inter-factor correlations within each model with at least two dimensions (i.e., 96% of models), we can see that virtually none of the factors belonging to the same model are negatively correlated while only a handful of (< 5%) approach orthogonality. At the same time, virtually no solutions are extremely highly correlated. Almost all factors are instead moderately associated with a mean correlation estimated at approximately 0.31. Panel B further shows that the number of latent dimensions is not systematically related to the inter-factor correlation; as we increase the item sample and estimate additional factors, such factors are not simply becoming more and more alike. The relatively moderate, mean factor correlation of 0.3 implies that most solutions *partially* capture the same underlying concept, albeit with considerable noise and heterogeneity. In sum, the solutions are different enough to warrant separate dimensions of *measured* social capital; they are similar enough, however, to reasonably assume that they are—at least partially—related to the *true essence* of the underlying concept.

Unfortunately, this very circumstance can be troublesome for research communities. If two scholars use slightly different but entirely reasonable sets of measurement items to assess the same concept, they can arrive at vastly different conclusions. Similarly, if the designers of the SCBC had fielded slightly differently phrased items, researchers would likely build upon (at least slightly) different results.

To illustrate this problem, we selected seven covariates from the SCBC that we correlat-

ed with each of the factors obtained in our simulation. These covariates—age, gender, education, religiosity, political ideology, and political knowledg—are prominently independent, dependent, or control variables in the empirical literature on social capital (Dinesen et al., 2020). We also added a seventh covariate, the language in which the survey was administered (English versus Spanish), expecting this feature to almost certainly fail to produce strong associations with social capital, no matter how we measure it.

Figure 5 presents probability density plots of pairwise correlations between the social capital factors and the above-listed covariates. While most associations are centered at zero, each plot provides ample room for scholars with different, perhaps slightly a-typical operationalizations of the concepts to legitimately claim social capital to be significantly negatively or positively associated with any of these covariates. These effects are particularly dramatic for age and (surprisingly!) survey language; here, mainstream scholarship might claim a substantial negative relationship while a vocal scholarly minority might (just as legitimately) come to the opposite conclusion.

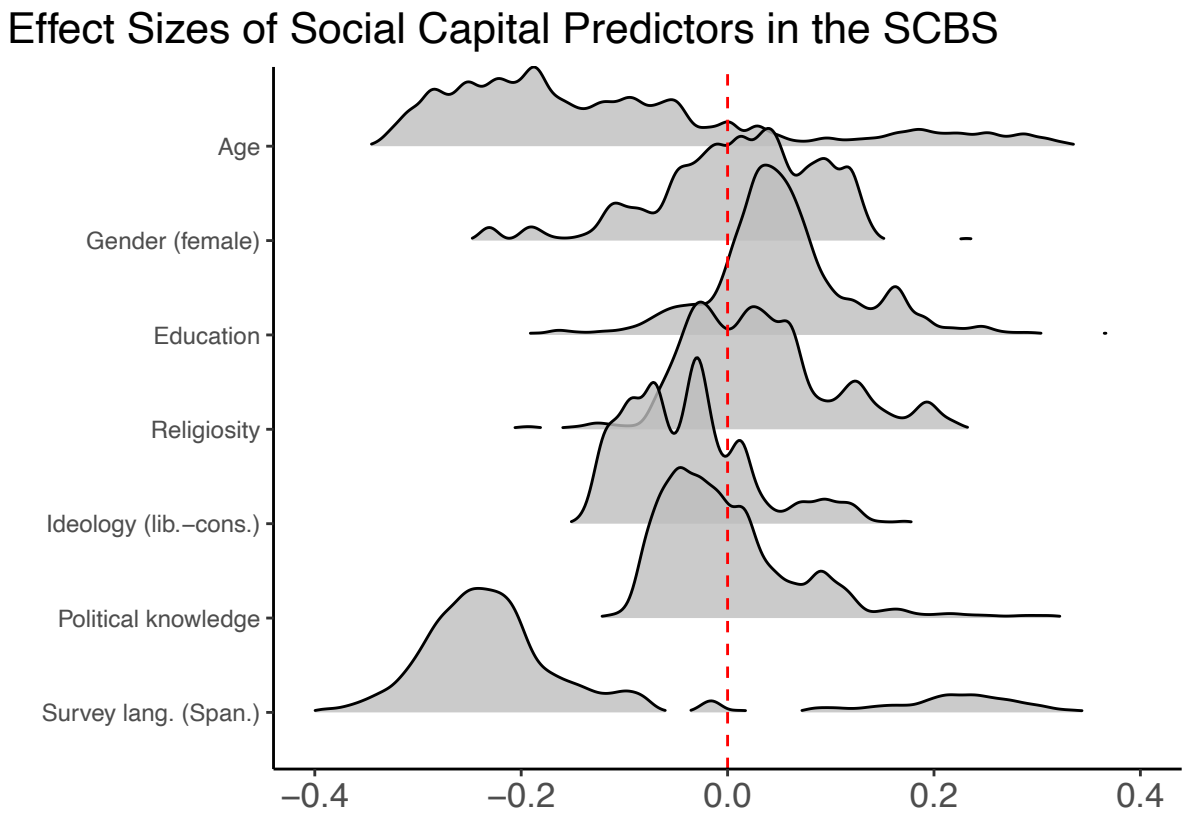

Figure 5. Covariate effects on social capital in the 2000 SCBS

CONCLUSION

Previous research has shown how different operationalizations of social science concepts can lead different scientists to reach other and sometimes opposite conclusions (Breznau

et al., 2022; Schröder & Yitzhaki, 2017; Carpentras & Quayle, 2023; Warncke et al., 2024). Some of the extant scholarly disagreements may be due to limited access to data. However, such disputes might ultimately be solved thanks to the big data revolution.

This article explored such an ideal-type scenario at multiple levels, using mathematical analysis, statistical simulations, and an empirical example. In all cases, we found strong evidence that differences imposed by different measurement operationalizations will not vanish even in the limit of infinite data. At the same time, relaxing the assumption of uni-dimensionality within the target concept does not necessarily improve reliability. We showed that extant measurement theories and methods cannot guarantee measurement convergence but that they are limited (or biased) by the data collection and selection process. At their best, measurement tools reduce dimensionality with a flavor of theory.

The problem at the core of numeric measurement (i.e., reducing a concept into a finite set of numbers) is surprisingly similar to answering a non-sensical question such as "What is the language spoken in Europe?". Even if we find a way to reduce the available *data* into a number, it does not mean that the same can be done for the *construct* itself.

Unfortunately, this problem is not just theoretical but has fundamental consequences for academic research and technological applications. For example, data-driven tools may attempt to maximize beneficial outcomes such as life satisfaction or social capital. However, such tools are not guaranteed to capture the concept itself (i.e., true life satisfaction) but only a specific empirical operationalization thereof.

Technology designed for universal improvement may, in fact, create hidden winners and losers. Consider the allocation of scarce resources as a central issue for society. We might create an ostensibly neutral, meritocratic allocation rule that makes specific funds for education or research only available to the top 5% according to a series of performance metrics. However, the results of Figure 3 show that 15% of the selected people would change even in near perfect conditions, simply by arbitrarily re-operationalizing this metric.

While the well-known issues of measurement error and sample bias might lend themselves to improvement through better measurement systems, the problem of operationalization is here to stay. If we want to advance scientific knowledge and technological innovation, we need to rethink how we measure constructs. Following the principle of Garbage-In Garbage-Out, if our measurement procedures remain deeply flawed, our tests and optimization approaches will always produce poor results.

However, we do not advocate for dropping the quantitative approach altogether. Instead, we need to catch up with technological progress and re-think the way in which we measure abstract concepts. Some new, alternative approaches are promising. For example, Attitude and Belief networks (Lueders et al., 2023; Boutyline & Vaisey, 2017) explore social complexity without reducing it into single numbers. Furthermore, they do not

attempt to build a latent variable but instead are uniquely based on human responses to stimuli.

Of course, the search for alternative measurement procedures implies more complexity, non-linear thinking, and conceptual flexibility. We may never be able to assign numeric scores to people or desired outcomes unequivocally. However, if our purpose is to measure complex constructs, then complexity should be embraced rather than reduced.

FUNDING: DC is grateful for support from the project "CoCi: Co-Evolving City Life", which was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement No. 833168.

CONFLICT OF INTEREST: The authors declare no conflict of interest.

DATA AVAILABILITY: https://github.com/just-a-normal-dino/measurement_problem_sim and https://github.com/pwarncke77/Big-data-delusion

REFERENCES

Bandalos, D. L. (2018). *Methodology in the social sciences. Measurement theory and applications for the social sciences*. New York, NY: Guilford Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., ... & Schonberg, T. (2020). Variabilityin the analysis of a single neuroimaging dataset by many teams. *Nature, 582*(7810), 84-88.

Boutyline, A. & Vaisey, S. (2017). Belief network analysis: A relational approach to understanding the structure of attitudes. *American journal of sociology*, *122*(5), 1371-1447.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... & Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119.

Caragliu, A., Del Bo, Ch., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of Urban Technology, 18*(2), 65-82.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage Publications.

Carpentras, D. & Quayle, M. (2023). The psychometric house-of-mirrors: the effect of measurement distortions on agent-based models' predictions. *International Jour-*

nal of Social Research Methodology, *26*(2), 215-231.

Carpentras, D. (2024). We urgently need a culture of multi-operationalization in psychological research. *Communications Psychology*, *2*(1), 32.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, *28*(3), 231-237.

Charitonidou, M. (2022). Urban scale digital twins in data-driven society: Challenging digital universalism in urban planning decision-making. *International Journal of Architectural Computing*, *20*(2), 238-253.

Costa Jr, P. T. & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, *13*(6), 653-665.

Courant, R., John, F., Blank, A. A., & Solomon, A. (1965). *Introduction to calculus and analysis* (Vol. 1). New York: Interscience Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

DeVellis, R. F. (2006). Classical test theory. *Medical care*, S50-S59.

Dinesen, P. T., Schaeffer, M., & Sønderskov, K. M. (2020). Ethnic diversity and social trust: A narrative and meta-analytical review. *Annual Review of Political Science, 23*, 441-465.

Duck-Mayr, J. & Montgomery, J. (2022). Ends against the middle: Measuring latent traits when opposites respond the same way for antithetical reasons. *Political Analysis*, *31*(4), 606-625.

Ghazal, T. M., Hasan, M. K., Alshurideh, M. T., Alzoubi, H. M., Ahmad, M., Akbar, S. S., ... & Akour, I. A. (2021). IoT for smart cities: Machine learning approaches in smart healthcare—A review. *Future Internet*, *13*(8), 218.

GIGO. (2024). Retrieved form https://it.wikipedia.org/wiki/Garbage_in,_garbage_out

Gligor, D. M., Pillai, K. G., & Golgeci, I. (2021). Theorizing the dark side of business-to-business relationships in the era of AI, big data, and blockchain. *Journal of Business Research*, *133*, 79-88.

Golino, H. F. & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One, 12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Golino, H., et al. (2021). Entropy fit indices: New fit measures for assessing the structure and dimensionality of multiple latent variables. *Multivariate Behavioral Research, 56*(6), 874-902.

Halperin, I. & Schwartz, L. (1952). *Introduction to the Theory of Distributions*. Toronto, ON: University of Toronto Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, *43*(3-4), 172-198.

Kean, J. & Reilly, J. (2014). Item response theory. *Handbook for clinical research: Design, statistics and implementation*, 195-198.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. New York, NY: Guilford Publications.

Kourtit, K. & Nijkamp, P. (2012). Smart cities in the innovation age. *Innovation: The European Journal of Social Science Research, 25*(2), 93-95.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Mineola, NY: Dover Publications.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (1989). *Foundations of Measurement Volume II: Geometrical, Threshold, and Probabilistic Representations*. Mineola, NY: Dover Publications.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (1990). *Foundations of Measurement Volume III: Representation, Axiomatization, and Invariance*. Mineola, NY: Dover Publications.

Krummel, T. M. (2019). The rise of wearable technology in health care. *JAMA Network Open*, *2*(2), e187672-e187672.

Labroo, A. A., Mukhopadhyay, A., & Dong, P. (2014). Not always the best medicine: Why frequent smiling can reduce wellbeing. *Journal of Experimental Social psychology*, *53*, 156-162.

Lai, C. S., Jia, Y., Dong, Z., Wang, D., Tao, Y., Lai, Q. H., ... & Lai, L. L. (2020). A review of technical standards for smart cities. *Clean Technologies*, *2*(3), 290-310.

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A.,...The Crowdsourcing Hypothesis Tests Collaboration.(2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin,146*, 451-479.

Li, X., Liu, H., Wang, W., Zheng, Y., Lv, H., & Lv, Z. (2022). Big data analysis of the internet of things in the digital twins of smart city based on deep learning. *Future Generation Computer Systems*, *128*, 167-177.

Luce, R. D. (1966). Two extensions of conjoint measurement. *Journal of Mathematical Psychology, 3*(2), 348-370.

Lueders, A., Carpentras, D., & Quayle, M. (2022). A Holistic View on Polarization: Attitudes, Emotions, and Partisanship as Elements of Social Identity Construction.

Retrieved from https://psyarxiv.com/apkzv/download?format=pdf

McNamara, M. E., Zisser, M., Beevers, C. G., & Shumake, J. (2022). Not just "big" data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour research and therapy, 153*, 104086.

Nakano, S. & Washizu, A. (2021). Will smart cities enhance the social capital of residents? The importance of smart neighborhood management. *Cities, 115,* 103244.

Nunnally, J. C. (1978). An overview of psychological measurement. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders: A Handbook* (pp. 97-146). Boston, MA: Springer.

O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE intelligent systems*, *28*(2), 96-99.

Pan, Y., Tian, Y., Liu, X., Gu, D., & Hua, G. (2016). Urban big data and the development of city intelligence. *Engineering*, *2*(2), 171-178.

Pedregosa, F., Varoquaux, Gael, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Pereira, G. V., Parycek, P., Falco, E., & Kleinhans, R. (2018). Smart governance in the context of smart cities: A literature review. *Information Polity*, *23*(2), 143-162.

Putnam, R. D. (1995). Bowling alone: Americas's declining social capital. *Journal of Democracy, 6*(1), 65-78. doi:10.1353/jod.1995.0002

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1990). Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests.

Roselli, D., Matthews, J., & Talagala, N. (2019, May). Managing bias in AI. *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 539-544). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3308560.3317590

Sanchez, F. & Sobolev, K. (2010). Nanotechnology in concrete–a review. *Construction and building materials*, *24*(11), 2060-2071.

Schröder, C. & Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, *92*, 337-358.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201-292.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Szarota, P. (2011). Smiling and happiness in cultural perspective. *Austral-Asian Journal of Cancer*, *10*(4), 277-282.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E.,...& Nosek, B. A. (2018). Many analysts, one dataset: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337-356.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*(2), 174-195.

Trencher, G. & Karvonen, A. (2020). Stretching "smart": Advancing health and well-being through the smart city agenda. In *Smart and Sustainable Cities?* (pp. 54-71). London: Routledge.

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *Journal of Theoretical and Philosophical Psychology*, *41*(1), 58.

Van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of item response theory*. New York: Springer.

Warncke, P., Searing, D.D. and Allen, N. (2024). Active, assertive, anointed, absconded? Testing claims about career politicians in the United Kingdom. *European Journal of Political Research*, 63(3), 1129-1154.

Webster, G. S. (1996). *A prehistory of Sardinia, 2300-500 BC* (No. 5). Sheffield: Sheffield Academic Press.

Wikipedia, https://en.wikipedia.org/wiki/Geographical_midpoint_of_Europe

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., & Vasilakos, A. V. (2024). GPT (Generative Pre-Trained Transformer) – A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE, 12*, 54608-54649. DOI: 10.1109/ACCESS.2024.3389497

Zhuang, Y. T., Wu, F., Chen, C., & Pan, Y. H. (2017). Challenges and opportunities: from big data to knowledge in AI 2.0. *Frontiers of Information Technology & Electronic Engineering*, *18*, 3-14.

BIOGRAPHICAL NOTE

Dino Carpentras is a postdoctoral researcher at the Centre for Computational Social Science, ETH Zurich. His research spans across multiple disciplines, including measurement theory, opinion dynamics and collective intelligence. Most of his work relies on agent-based or network-based modelling.

Philip Warncke is a political scientist who currently works as a postdoctoral fellow at Free University Berlin. His methodological research seeks to improve measurement properties of often ill-defined social science concepts including political ideology, careerism, and social trust. He also studies comparative political behavior in the United States and Europe.

ARTICLE HISTORY: Received 2024-07-14 / Accepted 2024-11-14