Research Article https://doi.org/10.12973/ijem.7.3.447

International Journal of Educational Methodology
Volume 7, Issue 3, 447 - 463.
ISSN: 2469-9632
https://www.ijem.com/

Data in the Educational and Social Sciences: It's Time for Some Respect

Henry Braun*
Boston College, USA

Received: January 1, 2021 ▪ Revised: April 10, 2021 ▪ Accepted: July 24, 2021

Abstract: This article introduces the concept of the carrying capacity of data (CCD), defined as an integrated, evaluative judgment of the credibility of specific data-based inferences, informed by quantitative and qualitative analyses, leavened by experience. The sequential process of evaluating the CCD is represented schematically by a framework that can guide data analysis and statistical inference, as well as pedagogy. Aspects of each phase are illustrated with examples. A key initial activity in empirical work is data scrutiny, comprising consideration of data provenance and characteristics, as well as data limitations in light of the context and purpose of the study. Relevant auxiliary information can contribute to evaluating the CCD, as can sensitivity analyses conducted at the modeling stage. It is argued that early courses in statistical methods, and the textbooks they rely on, typically give little emphasis to, or omit entirely, discussion of the importance of data scrutiny in scientific research. This inattention and lack of guided, practical experience leaves students unprepared for the real world of empirical studies. Instructors should both cultivate in their students a true respect for data and engage them in authentic empirical research involving real data, rather than the context-free data to which they are usually exposed.

Keywords: *Authentic data examples, carrying capacity of data, data analysis framework, quantifying uncertainty, teaching data analysis.*

To cite this article: Braun, H. (2021). Data in the educational and social sciences: It's time for some respect. *International Journal of Educational Methodology*, *7*(3), 447-463. https://doi.org/10.12973/ijem.7.3.447

Introduction

Empirical research, including data inspection and exploration, modeling, analysis and interpretation, plays a critical role in the educational and social sciences. It complements research of a more theoretical or conceptual nature by providing evidence to support or refute hypotheses and predictions. It can also generate surprises that spur new insights and models. Accordingly, graduate study in these domains generally includes courses in methodology comprising different combinations of quantitative, qualitative and mixed methods approaches. In the former, the emphasis typically is on developing families of statistical models (e.g., general linear models, multi-level regression models) with procedural recommendations, guides to the interpretation of the results and, occasionally, caveats and limitations. Qualitative methods courses describe different conceptual frameworks for the research, the procedures associated with each framework, along with guidelines for the conduct of the research (Shavelson & Towne, 2002).

Looking back on more than 45 years of theoretical and applied research, as well as many years of teaching, it strikes me as very problematic that even today there remains a deep disjuncture between actual practice and what is presented in most textbooks. In well-done, real world projects, investigators spend considerable time in scrutinizing, cleaning, and organizing the data. When the data have been collected as part of a project, data scrutiny includes a review of the data collection design, evaluation of fidelity of implementation of the design, various data checks, and documentation of any problems (e.g., departures from the implementation protocols, extent and patterns of missing data). Typically, the raw data is then transformed and organized into an analytic database – a process that may involve some combination of data exclusion, trimming, imputation and summarization.† When the project involves secondary analysis of data, similar scrutiny is conducted to the extent that the relevant information is available. The results are, or should be, foundational both to the conduct of the study and to the proper interpretation of the results.

* **Correspondence:**
Henry Braun, Boston College, Department of Measurement, Evaluation, Statistics, and Assessment; Chestnut Hill, USA. ✉ email

† This is a form of preprocessing that has implications for later inferential procedures (Blocker & Meng, 2013). This can be more problematic for secondary analysts who may not have access to the procedures employed. Data preprocessing also plays a role in considerations of data life cycles as discussed by Borgman (2019).

And yet … few first and second courses that I am aware of address this key aspect of applied research in any depth, so that researchers-in-training are given little background and few opportunities to practice the systematic inspection and evaluation of data. Furthermore, there is scant attention to how the implications for inference of this evaluation depend considerably on the context and purpose(s) of the study, as well as what relevant auxiliary information may be available. I characterize this neglect as a failure to cultivate in our students a proper respect for data. There are many culprits for this state of affairs: These issues and related procedures are rarely, if ever, addressed in textbooks while, at the same time, instructors are under pressure to cover an extensive and ever-expanding curriculum. Moreover, providing the necessary background and accompanying materials can be both burdensome and time-consuming.

This article offers an approach to addressing this disjuncture through the elucidation of the concept of the carrying capacity of a data set. In rough analogy with the notion from ecology of the carrying capacity of an environment, the carrying capacity of a data set (CCD) is defined as the inferential burden that can be supported by the data set. Note that when the carrying capacity of the environment is exceeded, then negative consequences often result (e.g., some degree of destabilization in the balance among species and/or environmental degradation). Similarly, when the CCD is exceeded the inferential chain from data to conclusions is weakened, thereby threatening (and even undermining) the credibility of those conclusions.

The CCD was introduced in a report (Braun, 1990), but remained largely dormant until the opportunity offered by an award prompted a return to and extension of the concept. By definition, understanding the CCD is crucial to the proper interpretation of the results of the analyses carried out on the data. Consequently, evaluating the CCD becomes an essential goal of any empirical research effort. But how is such an evaluation to be carried out? The article presents a CCD evaluation framework that comprises four distinct, ordered phases. A framework is simply a useful way of organizing and communicating a complex construct or a multi-step process. The framework offered here arose inductively as a distillation of hard-won experiences in a range of empirical investigations – experiences that too often included overlooked data problems, missteps in analysis and overly optimistic interpretations. Figure 1 displays a visual representation of the framework.

An important theme in CCD evaluation is the need to infuse data scrutiny and data-related considerations into the conduct of empirical studies. In recent years, the research literature has become more attentive to aspects of this issue, motivated perhaps by the rise of "Big Data". On the one hand, some authors have probed more deeply into the nature of data from a philosophical point of view (Leonelli, 2019). On the other, some have proposed frameworks to try to capture key stages of empirical research (Donoho, 2017; Keller et al., 2020). The CCD evaluation framework is complementary to the one offered by Keller et al. (2020). In particular, both frameworks make it clear that the concept of data quality does not have fixed meaning for a particular data set; rather, it depends on a number of factors including the intended use and the auxiliary information available.[‡] The contingent nature of data quality is perfectly analogous to the contingent nature of the validity of an assessment instrument: In the latter setting, validity is not regarded as an inherent property of the instrument itself but, rather, is dependent on purpose, context and use.

As described below, exploring and evaluating the CCD is a systematic way of revealing data quality in a particular setting. Donoho (2017, pp. 755-756) proposes a Greater Data Science (GDS) framework comprising six phases or divisions. Although there are a number of connections between CCD and some of the GDS phases, it is not the intent of this paper to explicate them. Suffice it to say that the GDS has greater scope, while CCD offers more detail on the careful examination of data and the implications for inference.

As will become clear, taking CCD seriously calls for adopting a more critical stance towards both standard data evaluations and conventional statistical analyses. Over the years, many statisticians have warned of the dangers of routine model building and the blithe acceptance of potentially problematic assumptions that has characterized much of empirical research.[§] Perhaps the most renowned was John Tukey. Remarkably, only 20 years after his introduction to statistics at the beginning of World War II, he already warned that statisticians were in danger of becoming increasingly irrelevant by ignoring the real problems of statistical practice.[**] In a seminal article Tukey (1962) argued:

Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole. Large parts of data analysis are incisive, laying bare indications which we could not perceive by simple and direct examination of the raw data… Some parts of data analysis … are allocation, in the sense that they guide us in the distribution of effort … in observation, experimentation, or analysis. Data analysis is a larger and more varied field than inference, or incisive procedures, or allocation (p. 3).

[‡] Both frameworks are consistent with the 'relational' perspective on data discussed by Leonelli (2019).

[§] By routine model building I mean the choice of a modelling strategy that involves little or no consideration of the relationship between theory and the variables included in the model, of data quality, and of the appropriateness of the strategy to both the data and the research questions posed.

[**] Donoho (2017) provides further examples of Tukey's prescience, both in his critique of academic (mathematical) statistics and in anticipating the utility of taking a more data-centric view of statistical practice.

The future of data analysis can involve great progress…Will it? That remains to us, to our willingness to take up the rocky road of real problems in preference to the smooth road of unreal assumptions, arbitrary criteria and abstract results without real attachments (p. 65).

Another was David Freedman who, in many articles and reports, cast a critical eye on many commonly used statistical models and the tenability of the assumptions that underlie their proper application (Freedman, 2010). His recommendation that practitioners expend some "shoe-leather" in thoroughly investigating the substantive context of the problem before embarking on model-based analyses is certainly consistent with the investigation of the CCD. In his landmark article on statistics and data science, Donoho (2017) calls out John Chambers, Jeff Wu, and Bill Cleveland as early proponents of a greater emphasis on (what we now refer to as) data analysis and less attention to classical statistical modeling and inference. With regard to conventional statistical modeling, Stark and Saltelli (2018) use the term cargo-cult statistics to refer to (and to deplore) "the ritualistic miming of statistics rather than conscientious practice" (p. 40).

Methodology

Carrying Capacity of Data

The carrying capacity of data (CCD) is an integrated, evaluative judgment of the credibility of specific data-based inferences, informed by quantitative and qualitative analyses, leavened by experience. The CCD is not a fixed quantity inherent in a dataset, as it depends crucially on the context and purpose of the analysis. Moreover, the CCD cannot be simply determined by inspection; rather, it emerges through an orderly series of investigations with due consideration of the patterns in the data that are revealed, as well as the nature of the intended inferences. By combining quantitative and qualitative considerations, the result is a nuanced evaluation of the utility – and limitations -- of the data at hand with respect to the substantive question(s) posed.

Figure 1 displays a schematic representation of the process by which the carrying capacity of a dataset can be evaluated. For convenience, this is denoted below as the CCD framework.

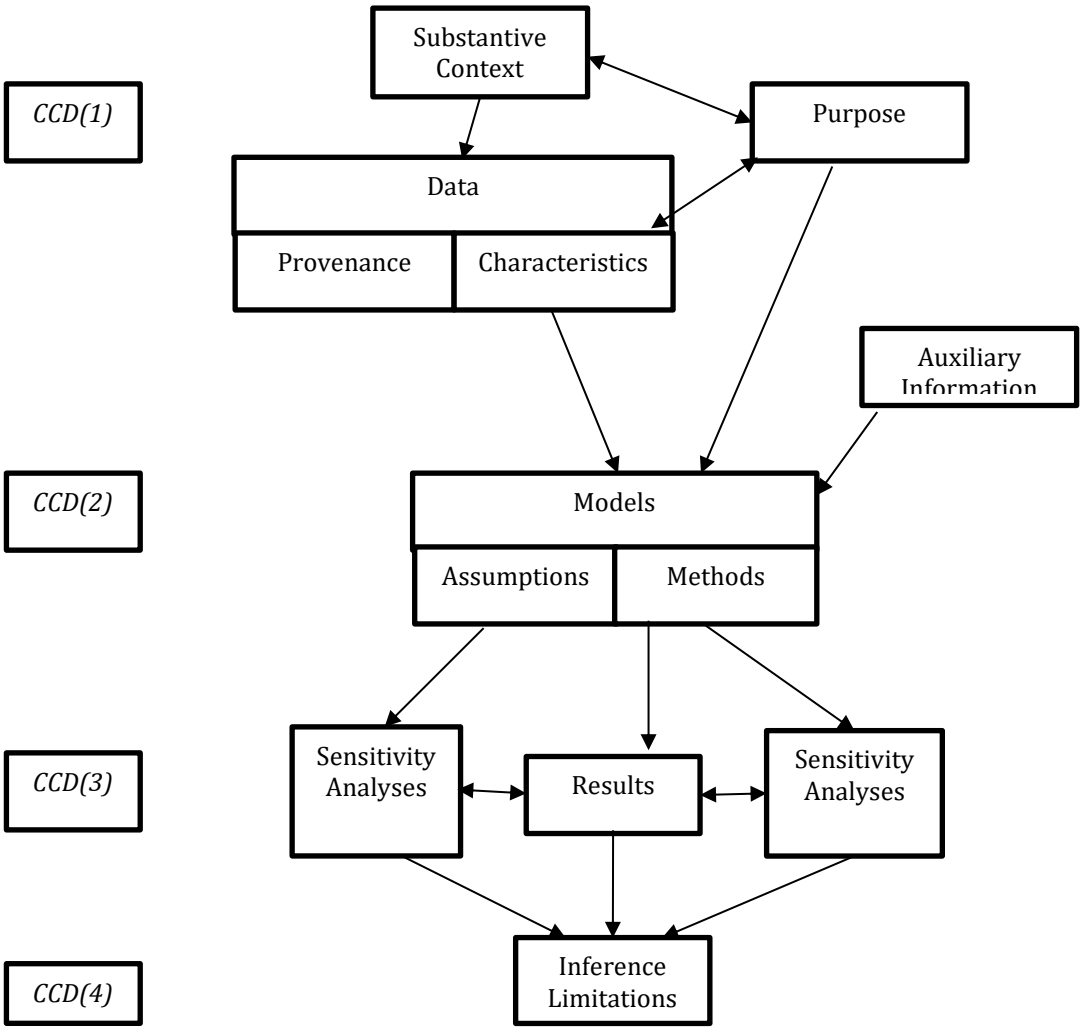

Figure 1. The Evaluation of the Carrying Capacity of Data: Schematic Framework

As depicted in Figure 1, the CCD framework comprises four phases that are briefly described in this section. In the next section, certain aspects of each phase will be illustrated with a number of examples. As the CCD can appear rather amorphous, that section also describes efforts to quantify the CCD in certain settings.

Phase 1

Purpose is framed within a particular substantive context and is often the result of a negotiation among stakeholders. In this phase the emphasis is on examining the available data in light of the problem context and the purposes of the analysis. (It bears mentioning that the CCD framework described here does not include considerations of alternative data collection designs). Given the design and the data at hand, questions are raised regarding

- the source(s) of the raw data,
- how that data was collected, reported, organized and (possibly) transformed,
- the grain size of the data,
- the treatment of missing data,
- how the analytic database was constructed,
- the likely power of conventional analyses.

Interrogating the source of the data can be revealing. For example, Wolf et al. (2020) show that estimates of the effectiveness of an intervention are systematically larger when the studies are done or commissioned by the developers than when they are done by independent investigators. As the responses to the questions above are collected and reviewed, concerns may arise and, in some cases, the purpose may have to be modified to address those concerns. Occasionally, analysis can indicate that the data are wholly inadequate to the purpose. For example, Bond and Lang (2019) show that survey data on individuals' levels of happiness (reported in ordered intervals) cannot reasonably be used to estimate two groups' relative average happiness – the usual estimand in such studies.

Finally, in preparation for the next phase, various descriptive statistics are produced so that relevant characteristics are available for study. These can include marginal and joint distributions, identification of outliers, along with data patterns that are evident from various graphical displays. Findings from this phase should be documented and reported appropriately.

With the emergence of "Big Data," raw data may undergo several stages of pre-processing before release to the analyst.[††] Some stages may be motivated by the need to summarize gigabytes of data, others to protect individual privacy (Leonelli, 2019). As an example, with the advent of computer-based testing, it is possible to record process data at the level of the individual keystroke or mouse movement. The enormous volume of data generated in a single testing session necessitates the construction of summary statistics (e.g., the number of keystrokes in each 20 second interval). These statistics can be used to monitor the testing process, complement existing item-level metadata, or identify anomalous respondent behavior. Deciding what summary statistics will prove most useful is an area of current research (Provasniak, 2021). Note that in the course of pre-processing, certain features of the data may be lost, or at least obscured, with implications for comprehensive data scrutiny.

Phase 2

Here the assumptions underlying different analysis strategies/models are evaluated based on the findings of Phase 1, as well as any auxiliary information that is available. Auxiliary information may introduce boundary conditions or otherwise cast doubt on the credibility of one or more of the assumptions. Alternatively, it may support the credibility of some assumptions. Ioannidis (2005) offers another example of the use of auxiliary information. He demonstrates that knowing that many independent studies are investigating the same phenomenon reduces the inferential value of any one study as a result of the problem of multiplicity. Ignoring multiplicity leads to too many false 'discoveries'.

Thoughtful use of displays, especially dynamic visualization, can enhance the CCD as anyone who has seen the You Tube videos of Hans Rosling can attest.[‡‡] Though this does not constitute auxiliary information in the usual sense, it is consistent with tenets of exploratory data analysis as developed by Tukey (1976).

In the era of Big Data, machine learning may be employed, either as a precursor or as an alternative, to classical statistical modeling. With machine learning, the emphasis is typically on discovering algorithms that produce useful predictions (according to a pre-determined metric), with less interest in being able to specify the pathway from input to output.[§§]

[††] These issues are by no means confined to the social and behavioral sciences. They may be even more problematic in "hard science" fields such as particle physics and helioseismology (P.B. Stark, personal communication, 6-21-2020). Combining quantitative and qualitative measures of uncertainty in environmental science under the NUSAP system is addressed in van der Sluijs et al. (2005). Similar issues in the world of business are discussed in Kennet and Redman (2019).

[‡‡] https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

[§§] Donoho (2017) points out this divergence in modeling strategies was already highlighted by Breiman (2001).

At the conclusion of this phase, a second determination of the CCD is made. It can signal potential trouble spots or weaknesses in the analysis strategy that may lead to supplementary analyses in the following phase.

Phase 3

In this phase, the main analyses are carried out, accompanied by various sensitivity analyses applied to both the assumptions and the methods. For example, when quasi-experimental designs are employed to estimate causal effects, analyses to evaluate sensitivity to hidden bias are called for (Rosenbaum, 2002). More generally, sensitivity analysis may involve applying different models, testing or weakening some assumptions, and so on. Typical assumptions to be examined are (i) the use of a particular parametric model for the objects of study, including error distributions; (ii) the mechanisms generating missing, truncated, or censored data; and (iii) the choice of prior distributions in a Bayesian analysis.

It is not uncommon for sensitivity analyses to suggest alternative approaches to be pursued. In any case, the outcome of this phase is not only a set of results related to purpose, but also a body of evidence regarding the CCD that comprises both quantitative and qualitative components. Ideally, this phase helps in generating more realistic estimates of the uncertainty to be attached to the results.

It is important to recognize that when data scrutiny (and/or data mining) leads to the choice of models to be investigated, then the usual interpretations of statistical output such as p-values may be seriously in error. Although strategies for dealing with post-selection inference appear in the literature (e.g., Berk et al., 2013; Tibshirani et al., 2016), unfortunately they are not much used in practice. As Berk et al. (2013) point out, the problem can be transformed into one related to simultaneous inference and, consequently, methods like control of the False Discovery Rate (Benjamini et al., 2009) may be particularly useful. Benjamini et al. (2019) provide further developments in this vein. For a more expansive and liberal view of the problem of selection, see Mayo and Cox (2010).

Phase 4

The final phase involves a review of the entire sequence of analyses, leading to an overall set of conclusions specific to the purpose of the study. In some cases, full or partial, answers are available, along with appropriate measures of uncertainty, as well as any limitations on the interpretation of the findings. In more extreme cases, the conclusion may be that the data at hand are insufficient to yield meaningful answers – but perhaps with some guidance as to the design of future studies.

Carrying out an analysis without seriously considering the CCD does not necessarily render the resulting interpretations and conclusions incorrect. On the other hand, it almost certainly means that the stated levels of uncertainty with respect to the conclusions underestimate the true uncertainty – often substantially so. More problematic, perhaps, is that the incorrect propagation of uncertainty can result in biased estimates, thereby increasing the overall mean squared error.[***] Such a problem arises in the area of international large-scale assessments such as TIMSS and PISA because of the sparse data designs employed. Successfully addressing the problem requires complex, statistical machinery that generates the so-called plausible values that underly the survey reports (von Davier & Sinharay, 2014).

In carrying out an evaluation of the CCD, the analyst should keep three critical points in mind. First, as indicated in Figure 1, the determination of the CCD evolves through the course of the investigation. In particular, it can change when new information becomes available or new analytic methods are brought to bear. Second, analysis of the CCD can help to pinpoint weak links in the inferential chain from data to conclusions - and even suggest alternative data analysis strategies, such as focusing initially on better-behaved subsets of the data. Finally, the determination of the CCD involves professional judgment – it cannot be reduced to the application of a set of algorithms.

The examples to be presented below, along with the accompanying discussion, strongly support the idea that students be alerted to the importance of what may be labeled forensic data analysis. Although the term has typically been applied to the detection of academic fraud, I suggest extending it to the conduct of all studies. Questions regarding the provenance of the data (where do they come from – and from whom), as well as the characteristics and deficiencies of the data, should be staples of any empirical study.

Indeed, broadly speaking, successive stages of a study can be represented as: forensic data analysis (FDA), exploratory data analysis (EDA), and confirmatory data analysis (CDA). FDA has been described just above. EDA represents an intermediate stage comprising the open-ended search for patterns and structure in the data, with heavy reliance on various visualization techniques. CDA is the inferential component of data analysis and the one that is emphasized in most standard courses on statistics. It comprises combinations of model estimation, hypothesis testing and (sometimes) sensitivity analyses. Of course, actual practice is rarely strictly linear, with multiple feedback loops the norm.

[***] I thank X-L Meng for making this point.

At this point, it may be useful to call out three examples that represent different levels of CCD. These – and others -- will be revisited in the next section.

- Low CCD. The 'Wall Charts' produced by the U.S. Department of Education during the 1980s was meant to provide states with comparative data that would help to inform policy decisions.
- Moderate CCD. The Tennessee STAR study was a randomized control trial intended to estimate the impact of smaller class sizes on student achievement
- High CCD. A comprehensive demographic survey in Malaysia was used to estimate the advantage of breastfeeding over bottle feeding on infant survival.

Findings

Scrutinizing the Data

The first step in evaluating the CCD is careful scrutiny of the data. As noted earlier, this is a key initial activity in any study not much dealt with in textbooks. Five examples should suffice to make the point.

(A) Organizations responsible for constructing and maintaining administrative and official databases expend considerable effort in achieving high levels of accuracy – due in part to their recognition of the extended life cycles of such data (Borgman, 2019). This can be particularly challenging when the data are obtained through multiple sources characterized by different levels of reliability and credibility.††† For example in a study of novice graduates of teacher preparation programs in one state, Braun et al. (2017) drew on multiple state administrative files to identify middle school teachers who had three or fewer years of classroom experience as the teacher of record. However, in a later phase of the study, a web survey of a subset of those teachers revealed that many were not in fact novice teachers according to the study's definition. The divergence between the information in the state databases and respondents' self-report added a degree of uncertainty beyond that provided by the conventional, model-based measures of variance.

(B) In other settings, there may be concerns regarding the accuracy of individuals' responses to surveys or other probes – particularly when the responses relate to sensitive issues such as salary, family circumstances, sexual orientation and the like. Goldhaber et al. (2019) offer a cautionary example involving teacher salaries in Washington State, with two databases yielding very different patterns. The authors highlight a conundrum:

> When data are not critical to any administrative process, documentation and reporting requirements are more likely to be lax, and the accuracy of such data should not be taken for granted. Where data *are* important to administrative processes … researchers should understand where there may be an incentive to misreport (p. 179).

(C) International large-scale assessments (ILSA) aggregate data from multiple jurisdictions, with each jurisdiction responsible for collecting data, either through governmental mechanisms or through third parties. Assuring data quality in this setting is very challenging. The Programme in International Assessment of Adult Competencies (PIAAC) is a case in point (Organisation for Economic Co-operation and Development [OECD], 2013). PIAAC is a household survey of adults ages 16-65 that conducts assessments of foundational skills and obtains rich background information. Data collected by trained staff is obtained through one-on-one interviews conducted with the aid of tablets. Data from the tablets is uploaded periodically to a central facility in the jurisdiction and then to the lead contractors. The auxiliary information available from the uploaded records facilitates checking some aspects of data quality. For example, a review of this information from the first round of data collection that took place in 2012 revealed that in the Moscow region of the Russian Federation some interviewers recorded impossibly large numbers of households visited and the data reported appeared to have been fabricated. Ultimately all data from the Moscow region was deleted from the official database.

(D) Meta-analysis is one area of research where there is an intense focus on data quality and relevance. The goal of meta-analysis is to summarize the results of multiple studies of the same phenomenon in order to determine if it is possible to reach a consensus regarding the nature of that phenomenon (Borenstein et al., 2009). In this context, each study is a datum. An initial scan of the literature may yield hundreds of studies. However, careful review of each study may result in only a small fraction retained for analysis. Although some studies are rejected because they are deemed not relevant, most are rejected because of deficiencies in the raw data, the documentation provided, the methodology employed or the reporting format. The credibility and utility of the meta-analysis depends crucially on the underlying assumptions, the exclusion rules employed, and the care with which the review process is conducted.‡‡‡

---

††† Many articles in a typical issue of the *Journal of Official Statistics* deal with examining and enhancing data quality – especially critical because these data often serve as input to secondary analyses with important policy implications. Statistics Canada and the US Census Bureau have been leaders in "data cleaning" and "data editing".

‡‡‡ It is possible that overly rigorous rejection of "deficient" studies could result in a final study set that yields a too-optimistic estimate of the effect. Berk and Freedman (2003) argue that most conventional meta-analyses are seriously flawed as they rely on assumptions that are not likely to be

(E) Although this article is concerned with quantitative data analysis, it bears mentioning that many aspects of CCD apply, *mutatis mutandis*, to qualitative data analysis as well. Most such studies collect evidence from a variety of sources such as one-on-one interviews, focus groups, administrative data and documents, and reports of various kinds. Before constructing a coherent narrative and reaching defensible conclusions, each datum should be evaluated with respect to both relevance and trustworthiness. This is done by careful consideration of the links of the datum to the research question(s), as well as by judgments about the credibility of the sources and triangulation among the different data. Reviews of the draft report by those who participated in the interviews or provided the data add assurance to the quality of the data and its interpretations.[§§§]

Data: Quality in Context

As noted earlier, the CCD must be evaluated in light of the purpose of the study. The CCD of the data may be quite adequate for one purpose but not for another, even if the FDA phase reveals no essential difficulties. Some examples illustrate the point.

(A) Retrospective ascertainment data arise when the time of occurrence of an event is not known until a second event occurs. A classic example is infection with HIV due to transfusion with blood containing the virus. The point of infection is not recorded until the individual develops AIDS and it is subsequently determined, with reasonable certainty, that the cause was the transfusion. One question is whether this sort of data can be used to estimate the AIDS incubation time. This was addressed by Kalbfleisch and Lawless (1989) and they were able to determine the limitations of the data in answering the question. A similar analysis was undertaken by Braun (1989) employing data on bid-rigging convictions to estimate the probability that such crimes would be uncovered and the perpetrators convicted. It was found that such estimates could be obtained but were very sensitive to model assumptions – assumptions that could not be independently evaluated.

(B) In 1984, the U.S. Department of Education released the "Wall Chart," a compendium of information organized by state. The chart comprised three panels: the first panel contained various performance indicators, the second resource inputs, and the third population characteristics. As noted by Ginsberg et al. (1988), this was the first time that extensive state-level data, including performance data, were available in a format that facilitated comparisons among states. Indeed, the stated intent was to encourage such comparisons so that state leaders and other stakeholders could identify "what works" and make appropriate policy adjustments.

(C) Although Ginsberg et al. (1988) were quite supportive of the effort, there was considerable skepticism that such data, however analyzed, could yield useful policy evidence as to which states had the more effective education systems (Wainer, 1986a). For example, a key performance indicator in the Wall Chart was related to a state's average score on one of two national, college admissions tests (ACT or SAT).[****] States were ranked on the basis of the changes in average test scores (on the ACT or the SAT) over a ten-year period. It had long been recognized that states' average scores on these tests were correlated with the percentages of students in the high school cohort taking the test and there had been different attempts to adjust these scores for differential selection bias (e.g., Steelman & Powell, 1985). Others argued, however, that such adjustments depended on assumptions that could not be empirically verified and, unsurprisingly, the results were rather sensitive to the method employed (Holland & Wainer, 1990; Wainer, 1986b). Notably, Holland and Wainer (1990) drew on auxiliary information to cast doubt on a key assumption of the adjustment methods. Among other things, this example illustrates how there can be honest debates regarding the CCD, especially with respect to informing policy decisions.

(D) In a study of the feasibility of evaluating the quality of teacher preparation programs using novice teacher value-added scores, Braun et al. (2017) argued that the data available were inadequate to the task; that is, the CCD was not sufficient for the purpose. Among the concerns cited were: (i) the small number of graduates in different programs (e.g., middle school mathematics); (ii) novice teacher efficacy is associated with the extent and quality of the induction/mentoring program at the school-of-placement, but the information on such programs was not readily available; (iii) the proportions of graduates that obtain employment in the state's regular public and charter schools (and so are included in the state's administrative databases) varies substantially across programs. Further, in more selective programs many stronger graduates take jobs out of state and/or choose to work in private or parochial schools, while students in less selective programs are more likely to work in-state or fail to find jobs; (iv) 95% confidence intervals for program-level mean value-added scores employing model-based estimates of variance were so broad that even programs widely separated in rank had overlapping intervals. Braun et al. (2017) concluded that the combination of high variance, substantial bias and questionable data quality rendered this approach to program evaluation unworkable.

satisfied in practice and, moreover, their conclusions are very sensitive to those assumptions. In our terms, the CCD can be rather weak. For a more flexible approach to meta-analysis utilizing Empirical Bayes models, see Hedges (1988).

[§§§] For an extended example, see Hargreaves and Braun (2012).

[****] States were divided into two groups, depending on whether their students were more likely to take the ACT or the SAT.

(E) Data from ILSA are used for many different purposes. Table 1 is adapted from a report on two workshops carried out under the auspices of the U.S. National Academy of Education (Singer et al., 2018). The figure displays seven such purposes along with the editors' judgment of the general suitability of ILSA data for each purpose. In effect, these are judgments of the CCD of ILSA data for these purposes, ranging from providing relevant information about jurisdictions' education systems to supporting causal inferences regarding specific education or social policies.

These examples illustrate how conducting a thorough analysis of the adequacy of the data for the purpose of the study can help to set appropriate limits on the kinds of conclusions that can be drawn, as well as the credibility to be attached to those conclusions. Such analyses typically involve technical matters and disagreements among experts are not unusual. Indeed, a lack of consensus should sound a note of caution.

Table 1. The Seven Major Uses of ILSAs—How Well Can ILSAs Achieve These Goals?

	Uses	**How well can ILSAs achieve this goal?**
1	To be ***transparent*** about the condition of a nation's education system	Outstanding
2	To disturb complacency and ***spur education reforms***.	Outstanding
3	**To *describe* and *compare*** student achievement and contextual factors (e.g., policies, student characteristics) across nations.	Yes, with some caveats
4	**To *track changes over time*** in student achievement, contextual factors, and their mutual relationships, within and across nations.	Yes, with some caveats
5	To ***create de facto international benchmarking***, by identifying top-performing nations and jurisdictions, or those making unusually large gains, and learning from their practices.	Challenging, with many caveats
6	To ***evaluate*** the effectiveness of curricula, instructional strategies, and education policies.	Only with extreme caution
7	To ***explore causal relationships*** between contextual factors (demographic, social, economic and educational variables) and student achievement.	Generally impossible

Auxiliary Information

Judgments of CCD can be influenced by relevant auxiliary information. Such information may be obtained from related studies, other databases or from individuals with knowledge of the data or the context. Of course, the impact can be either positive or negative.

(A) The Tennessee STAR experiment (Nye et al., 2001) was a field RCT undertaken to examine the impact of both class size and class staffing on children's academic progress as measured by standardized test scores. For the most part, analyses indicated that smaller class sizes led to improved test scores, especially for more disadvantaged students. Since the initial reports appeared, questions were raised about how well the original design had been realized: for various reasons, randomized field trials are never perfectly executed. Chetty et al. (2011) used data from tax records linked to students' families to support the assumption of equivalence of treatment groups. They found no systematic differences in the family incomes of the children in the various treatment groups. Accordingly, judgment of the CCD of the data for inferences about treatment effects was enhanced.

(B) The U.S. National Assessment of Educational Progress (NAEP) is a large-scale assessment survey that collects nationally representative achievement data on fourth and eighth graders in reading and mathematics.†††† The results are used primarily to make comparisons of performance distributions among states and among sub-populations defined by combinations of individual characteristics such as gender and race/ethnicity. Important outcomes are estimates of achievement gaps between (say) White students and Black or Hispanic students – nationally or by state. The sampling of schools and students within schools is conducted with high fidelity so that generalizability should be high. On the other hand, there are no stakes for students, leaving persistent questions regarding their motivation. Obviously, differential motivation and effort expended across states and/or sub-populations can lead to biased results. Previous small-scale studies yielded conflicting results. Braun et al. (2011) conducted an RCT with nearly 9000 students drawn from a broad range of schools in seven states. The focus was the NAEP 8th grade reading assessment. Students were administered a "NAEP-like" form under conditions that mimicked those of the formal assessment. The experiment comprised three treatment arms: a control, a fixed incentive and a contingent incentive. Pairwise comparisons of score distributions between treatment arms for many subpopulations yielded differences that were statistically significant and substantively meaningful – especially between the control and the contingent incentives, with lower performance in the control condition. One conclusion was that current estimates of achievement gaps are likely underestimating the true achievement gaps, thereby lowering confidence in the CCD of NAEP for such inferences.

†††† NAEP tests 12th graders on a quadrennial cycle. It also tests other subjects in different grades as budgets permit.

There is evidently great value in bringing auxiliary information to bear though it often takes some imagination to identify potential sources. At the same time, the evaluation of the quality and relevance of such information should be carried out with the same rigor as for the original data. When the issue is sufficiently important and the degree of uncertainty too great, special studies can be conducted to produce additional, relevant information.

Sensitivity Analysis

In evaluating the CCD, sensitivity analysis denotes a range of strategies for ascertaining how robust are the intended inferences to the assumptions and models from which the inferences are derived. If the inferences appear strongly dependent on specific assumptions and models, then a judgment regarding the CCD is markedly diminished. In classical statistics, non-parametric methods were introduced for just this reason: Properties of non-parametric estimates (say, of location) or of non-parametric test statistics were at most weakly dependent on the underlying distribution of the data. The trade-off was lower relative efficiency for a given distribution in comparison to parametric methods chosen to optimize performance for that distribution. This, in turn, led to the development of robust-efficient methods that maintained high levels of relative efficiency for a broad range of distributions. Estimates of location and estimates of regression models are perhaps the best-known examples (Maronna, 2018). In the present context, we can say that these methodological developments enhanced the CCD of the data for certain types of inferences.

Of course, there are many different approaches to sensitivity analysis, too many to address in the present paper.[‡‡‡‡] Three examples follow.

(A) A randomized control trial in Tennessee of a pre-K intervention found effects at the end of the school year, but that overall the advantage had dissipated by the third grade. Pearman II et al. (2020) investigated whether there were combinations of factors that resulted in sustaining the early advantage. Of the original student sample of 1240, more than a third (434) were missing one or more data components. The authors chose to work with the complete data sample and to conduct sensitivity analyses to determined how robust the results were to that choice, as well as other decisions made during the course of the analysis. With respect to the latter, a number of alternative strategies were implemented. The results displayed only minor deviations from those derived in the original analysis. To assess the sensitivity to using only the complete data sample, they carried out imputations of the baseline covariates and repeated the analysis for the full sample. Again, only minor deviations were observed, hence enhancing the CCD for the study. Ideally, different imputation models would be implemented to more fully examine the robustness of the results. In cases where substantively meaningful differences are found through sensitivity analysis, it is an open question on how to report the uncertainty in the results. Resorting to Bayesian modeling may be called for.

(B) Another important area of application is the analysis of data from quasi-experimental designs (QED), where hidden bias is an ever-present concern. The question is whether the observed covariates used to control for self-selection, (e.g., through regression-based adjustment, propensity score weighting, etc.) have fully removed the bias induced by self-selection and other factors (Dearing & Zachrisson, 2019; Rosenbaum, 2002).

Conventional analysis employs measured covariates to take account of pre-treatment differences in the treatment and control groups. However, interpreting the coefficient of the treatment indicator as an (approximately) unbiased estimate of the true treatment effect depends on the assumption that the regression-based adjustment has accounted for essentially all the selection bias. An approach to sensitivity analysis proposed by Rosenbaum and Rubin (1983) is to embed the current best model in a richer model that includes an unobserved, individual-level variable (denoted U) that is posited to capture the remaining hidden bias. One then makes a series of assumptions about the relationship of U to the probability of selection into the treatment or control groups, and to the outcome variable. For each pair of assumptions about U, a new estimate of the treatment effect is obtained.[§§§§] The assumptions are directional in the sense that they are intended to produce lower estimates of the treatment effect.

As the assumptions are varied systematically, they generate a response surface of estimated treatment effects. Figure 2 displays two such response surfaces.[*****] Figure 2a illustrates the situation in which the conventional estimate is very sensitive to departures from the (naïve) assumption that there is no hidden bias; that is, the estimated treatment effects decline precipitously with relatively small deviations from that assumption. By contrast, Figure 2b illustrates the situation in which the conventional estimate is relatively robust and declines rather slowly as the assumptions about U become more extreme. Note that deciding whether the response

‡‡‡‡ For a somewhat different perspective, employing global sensitivity analysis, see Becker et al. (2014). Another approach, employing the coefficient of proportionality has been proposed by Oster (2019).

§§§§ These estimates are also adjusted for the observed covariates.

***** These figures are taken from Diaconu (2012).

surface more resembles Figure 2a or Figure 2b is often a matter of judgment and will depend on the context, the data available and the magnitudes of the estimated coefficients in the base model.

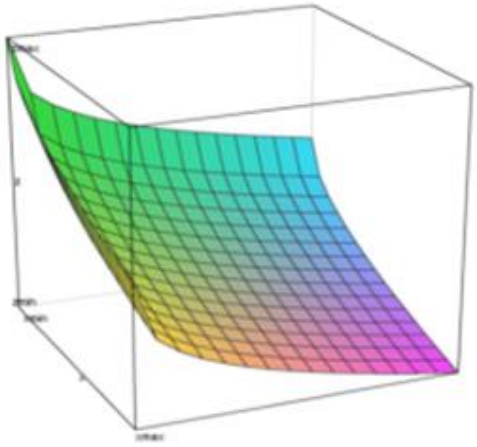

Figure 2a. Example of a steep response surface

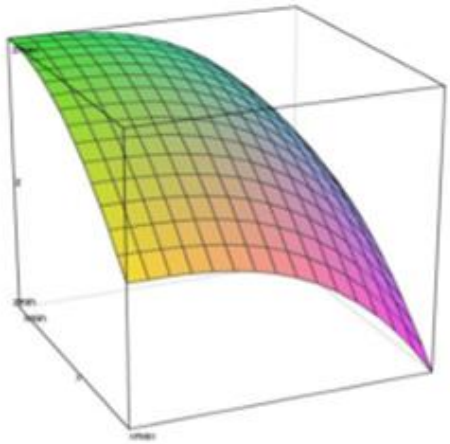

Figure 2b. Example of a shallow surface

Figure 2. Response Surfaces for Sensitivity Analyses

Many different variations and extensions of this approach can be found in the literature (Carnegie et al., 2016; Rosenbaum, 1989, 2002). An et al. (2018) carried out a sensitivity analysis for an elementary school intervention that focuses on addressing non-academic barriers to learning (Walsh et al., 2014). The outcomes of interest were the students' performances on state-level, middle-school assessments of mathematics and English/Language arts. A conventional analysis employing both propensity score matching and regression adjustment yielded estimates of the treatment effects that were positive and statistically significant. Effect sizes were modest to moderate.

The sensitivity analyses indicated that the estimated effects and effect sizes were reasonably robust to substantial hidden bias. Indeed, the derived estimated effects fell within the one-sided 90% confidence intervals from the standard model. The authors noted, however, that sensitivity analysis addresses only one of the six main threats to the validity of causal estimates of treatment effects from observational studies (Reardon & Raudenbush, 2009). For example, one threat is "no interference between units"; that is, the outcome for each unit is independent of the outcomes of the other units. That assumption is rather problematic when the units are students nested within classrooms nested within schools. Unfortunately, estimating the impact of departures from this assumption is a non-trivial exercise. This underscores the importance, if not the necessity, of drawing on auxiliary information in evaluating CCD.

(C) Sensitivity analysis can sometimes take advantage of auxiliary information. A case in point is the Stanford Education Data Archive, an ambitious project to link average test scores in English/Language arts and mathematics for grades 3 through 8 across all school districts in the U.S. Through a complex sequence of transformations, state test score scales were linked to the (common) NAEP scale. The construction of such a national reporting scale rests on a number of non-trivial assumptions. In addition to testing some of the assumptions directly through a form of cross-validation, Reardon et al. (2021) describe a complementary approach that employs test score data generated by a private test score vendor for about twenty percent of the school districts in the country. District-level, precision-adjusted correlations between the linked test score averages and the vendor test score averages ranged from 0.85 to 0.95, averaging 0.93. These and other findings indicated that the assumptions collectively did not induce unacceptable levels of bias, lending credibility to the linked scale. For further discussion of these issues, including some criticisms, see McCaffrey and Culpepper (2021).

Quantifying the CCD

To this point, the discussion has focused on raising concerns about how failure to carefully scrutinize the data and to consider the plausibility of model assumptions in light of both data limitations and contextual constraints can lead to unwarranted confidence in the results. A natural question is whether it is possible to provide sharper guidance in form of quantitative bounds. The answer is yes but, as one might expect, much depends on the particular context. Some examples follow.

(A) Over the last fifteen years there has been much interest in using longitudinal student test score records to estimate a teacher's relative contribution to student growth. One family of models, termed Value-added Models (VAM), were formally introduced by Sanders et al. (1997). Rather complex calculations yielded a value-added "score" for

each teacher/classroom or grade or school.†††† Because VAM incorporate test scores from prior years as predictors, they are generally only computed for teachers of English/Language arts or Mathematics in grades four to eight. Although there has been much criticism of the appropriateness of employing VAM scores for teacher evaluations (Darling-Hammond et al., 2012), value-added analyses remain a popular research tool. One technical concern focuses on the random effects model for VAM favored by statisticians. The issue is related to the possibility of a non-trivial correlation between teacher random effects and the random error component, resulting in biased estimates. As noted earlier, bias is particularly problematic not only because it is hard to quantify, but also because it can affect both accuracy and measures of uncertainty. For a particular family of VAM models, Lockwood and McCaffrey (2007) derived an exact expression for the bias and showed that the bias tended to zero quite rapidly as the number of prior test scores employed in the model increased.

(B) Another source of bias in VAM is the non-random sorting of students and teachers. Using data from North Carolina, Rothstein (2009) estimated the bias in teacher estimates for various model specifications, with varying assumptions about how selection depends on both observables and non-observables. In many cases the estimated bias was large enough to raise serious concerns about the use of the estimates in teacher evaluation.‡‡‡‡‡

(C) The problem of estimation bias in QED is ever-present. Montgomery et al. (1986) analyzed data from a demographic survey conducted in Malaysia. Interest centered on estimating differential infant survival depending on whether the child was breastfed or bottle-fed. The database contained extensive background information on the mother-child dyad. For the child it recorded year of birth, birthweight (dichotomous: low or high), ethnicity, breastfeeding history, and whether the child had survived to one month and then to one year. Conventional logistic regression analyses yielded a very substantial advantage to breastfeeding for one month survival and a smaller, but still substantively and statistically significant, advantage for one year survival conditional on one month survival. Montgomery et al. (1986) then carried out a sensitivity analysis for each time point using an extension of the method proposed by Rosenbaum and Rubin (1983). They found that the conventional estimates were robust to hidden bias. Finally, taking advantage of the longitudinal aspect of the data, they were able to obtain specific estimates of the treatment effects (at one month and at one year) from a logistic multiple regression model that incorporated a variable representing unobserved, individual-level heterogeneity.§§§§§

(D) It is rare that a designed experiment—a randomized control trial—is conducted on a set of units that have been randomly sampled from the target population. One consequence is that although treatment effect estimates may have high internal validity, their external validity, or generalizability, is in doubt. Over the years, there have been a number of attempts to quantify the degree of generalizability. They rely on some measure of the quality of the match on various relevant characteristics between the set of experimental units and the target population. This literature is reviewed by Tipton and Olsen (2018). Tipton (2014) developed a generalizability index that constitutes a useful, quantitative guide to how likely the obtained experimental estimate, after post-stratification, would be close to an estimate based on a random sample from the population. The index is the product of three factors and ranges from 0 to 1, with values near 0 indicating low generalizability and values near 1 indicating high generalizability.

(E) In his study of inference problems with Big Data, Meng (2018) considers, among other issues, the bias in estimating the population mean from a sample mean, where the sample is not necessarily a probability sample. Coincidentally, he derives a formula that is also the product of three factors, representing data quality, data quantity, and problem difficulty. Using the bias estimate, he is able to quantitatively compare the utilities of different sampling procedures – in other words, their CCD for a particular inferential problem. Meng emphasizes that data quality (and utility) is contingent on the context and purpose of the analysis.

Discussion

One theme of this article is that in order to arrive at defensible inferences accompanied by appropriate estimates of uncertainty, an essential first step is the careful and thorough investigation of the characteristics of the data. This should include considerations in the pre-processing of data, an activity that is increasingly popular in this era of Big Data. More broadly, the CCD framework offers a general roadmap for how to implement the stages of forensic, exploratory and confirmatory data analysis. Equally important, the framework is sufficiently rich to support various pedagogical strategies to expose students early on to the real world of applied statistics and data science.

The issue of pedagogy is certainly not new. Nearly 60 years ago Tukey (1962) expressed concern that students of statistics were not being given sufficient exposure to real data and the opportunity to apprentice with working data analysts. Singer and Willett (1990) decried the near ubiquitous use of artificial data in statistics courses and offered

†††† Since then, many other models have been proposed to estimate teachers' contributions to student learning (Braun & Wainer, 2007).

‡‡‡‡‡ Although it is beyond the scope of this paper, there are important questions regarding the relationship between data limitations, model misspecification and the use of the results for various purposes (Braun, 2015; Harris, 2009).

§§§§§ To obtain these estimates, it was necessary to assume that certain higher-order interactions were time-invariant. Those assumptions deserve further examination.

suggestions on how to bring real data into the classroom in order to afford students more authentic experiences with scientific work. With the growth of internet-based resources comprising both data sets and instructional modules, the situation has certainly improved over the last three decades.

Nonetheless, most students remain woefully underprepared for the real world of data analysis and, in particular, they lack an understanding of the essential role of data forensics, broadly defined. To this point, Donoho (2017) asserted that "Data Gathering, Preparation and Exploration" (the first phase of his GDS framework) was not only more important than "Data Modeling" (phase five of his framework), but typically consumed more effort and resources.

What is to be done? In individual courses, instructors can help students recognize the importance of conducting a thorough scrutiny of the data before plunging into modeling and confirmatory analysis – and how the results, along with findings from sensitivity analyses, can yield a refined estimate of the CCD. Instructors should be cultivating in their students an appreciation for the utility of thoroughly evaluating the CCD, as well as a healthy skepticism towards conventional interpretations of model-based inferences.

An initial step in that direction would be replacing or supplementing "toy data sets" with carefully curated examples that can be employed to illustrate both forensic and exploratory data analysis. Ideally, some of the examples would be culled from the instructor's own work so that relevant contextual information could be provided. A complementary strategy is to have students pose a problem that requires them to collect some data either directly or indirectly by extraction from an existing database. It is regrettable that the empirical examples found in most textbooks (or in their online supplements) present data that are either fabricated or "real," but with no or limited background information. Rarely are data presented with sufficient detail to permit meaningful scrutiny and consideration of the implications for inference. When large data sets are employed for pedagogical purposes, it is usually to illustrate the application of a software module, with the emphasis on procedures and straightforward interpretations of output. There is neither discussion of the relationship of the characteristics of the data to the credibility of the desired inferences nor of the sensitivity of those inferences to assumptions about the data or those underlying the statistical models employed.

An important benefit of authentic, data-centered instruction is that it provides a solid basis for developing students' understanding of common pitfalls, ranging from data collection problems and coding errors to a failure to appreciate the inherent limitations of the data for the problem at hand. Experience with real data and the inevitable vicissitudes of actual data analysis and modeling can make courses in quantitative analysis and data science both more engaging and more memorable! In this regard, the CCD framework can guide the design of a sequence of modules that would help prepare students for successful professional practice. The modules could be embedded in different courses with the goal of introducing students over the course of a program of study (masters or doctoral) to the importance of early investments in ascertaining data quality, along with increasingly sophisticated approaches to evaluating the CCD in a range of settings. This would be analogous to the Tuning Project, which involves the redesign of undergraduate programs, with the goal of strengthening curricular coherence in the majors in order to enhance student learning (Jankowski & Marshall, 2017).[******]

Conclusions

This article introduced the concept of the carrying capacity of a data set (CCD) and presented a rationale and framework for the thorough evaluation of the CCD as an integral part of any empirical study. Clearly, having CCD evaluation as standard practice remains very much aspirational. Although most studies do carry out a number of the evaluation components, few conduct them all. Aside from investigators not recognizing the need for some components, there are often constraints related to time and cost that mitigate against a complete CCD evaluation. It is not uncommon for investigators to fail to commit sufficient time or funds for a thorough forensic data analysis.

More problematic is the fact that even when limitations to the CCD are known, they are exceeded due to external forces. These include incentives to generate 'statistically significant' findings and the pressure to use extant data to answer questions for which the data are not appropriate. One manifestation is that known defects in the data are ignored when deriving measures of uncertainty for estimated parameters (e.g., ignoring bias) and, subsequently, when interpreting the results. To paraphrase one colleague's comment, in the current climate, exceeding the CCD of one's data may be the only way to thrive – if not survive.

Similarly, there are obstacles to incorporating aspects of CCD evaluation into courses. Devoting more time to data forensics, for example, would necessitate making difficult choices on allocating less time to more standard content. It would also require instructors to invest resources in identifying appropriate data sets and devising the accompanying instructional modules. In most tertiary institutions, especially those with a focus on research, the rewards for major modifications to existing courses are meager at best.

With these considerations in mind, I argue that we must change the circumstances under which we teach and publish. With respect to the former, we could begin by building an open source archive of curated data sets that include sufficient background information for at least a modest exercise in data forensics. If the data sets are sufficiently rich, other aspects

[******] For a related proposal on training of specialists in educational measurement, see Russell et al. (2019).

of CCD evaluation could be carried out (e.g., sensitivity analyses). Over time, the archive would grow to support methodological instruction with different substantive foci and, as a result, reduce the time required to integrate the material into a course.

With respect to the latter, journals must set more demanding standards for the publication of empirical analyses. There are precedents for such changes. In clinical research, in comparison to the standards in place thirty years ago, journals now require more sophisticated statistical analyses, as well as more thorough discussion of both the rationale for the choice of models and the limitations of the analyses presented. The ongoing controversy regarding the overuse/misuse of p-values has prompted changes in the publication guidelines for some journals. Although, such changes do not occur overnight, a five- to ten-year horizon is not impractical.

Limitations

In the evaluation of the CCD, the focus in this article has been on quantitative methods and the types of data to which they are typically applied. It was noted in passing that many of the considerations apply to qualitative methods of analysis. However, that branch of empirical analysis is much less familiar to the author, who was therefore reluctant to venture along that branch. Surely, it would be very useful for others to extend the CCD framework to accommodate such methods and, ultimately, to mixed methods as well. Of course, the evaluative judgments rendered at each stage of an investigation, although informed by quantitative analyses, are essentially (and necessarily) qualitative in character.

Another obvious limitation is that this article presents a personal view of a general strategy for statistical analysis and inference, along with its implications for both practice and pedagogy. Undoubtedly, others expert in these areas would offer somewhat different perspectives and emphases. Nonetheless, the hope is that the article will at the least stimulate a healthy debate on how to improve the quality of empirical research and related pedagogy in the educational and social sciences, both now and in the future.

Acknowledgments

The author would like to thank Y. Benjamini, R. Brennan, S. Konstantopoulos, JR Lockwood, L. Ludlow, X-L Meng, M. Russell, P.B. Stark, and several anonymous reviewers for helpful comments and suggestions. E. Angel and K. Borowiec provided research assistance.

References

An, C., Braun, H. I., & Walsh, M. E. (2018). Examining estimates of intervention effectiveness using sensitivity analysis. Educational Measurement: Issues and Practice, 27(2), 45-53. https://doi.org/10.1111/emip.12176

Becker, W., Paruolo, P., & Saltelli, A. (2014). Exploring Hoover and Perez's experimental design using global sensitivity analysis. arXiv. https://arxiv.org/pdf/1401.5617.pdf

Benjamini, Y., Hechtlinger, Y., & Stark, P. B. (2019, June 2). Confidence intervals for selected parameters. arXiv. https://arxiv.org/pdf/1906.00505.pdf

Benjamini, Y., Heller, R., & Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A, 367*, 4255-4271. https://doi.org/10.1098/rsta.2009.0127

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics, 41*(2), 802-837. https://doi.org/10.1214/12-AOS1077

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds.), *Law, punishment and social control: Essays in Honor of Sheldon Messinger* (2nd ed., pp. 235-254). Aldine de Gruyer.

Blocker, A. W., & Meng, X. L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernouilli, 19*(4), 1176–1211. https://doi.org/10.3150/13-BEJSP16

Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy, 127*(4), 1629-1640. https://doi.org/10.1086/701679

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.9a36bdb6

Braun, H. I. (1989, August 6-11). *Analysis of retrospective ascertainment data in a legal setting* [Paper presentation]. 1989 Annual Meeting of the American Statistical Association, American Statistical Association, Washington, DC, United States.

Braun, H. I. (1990). *Data in the social sciences: It's time for some respect* [Unpublished report]. Educational Testing Service.

Braun, H. I. (2015). The value in value-added depends on the ecology. *Educational Researcher, 44*(2), 127-131. https://doi.org/10.3102%2F0013189X15576341

Braun, H. I., Jenkins, F., & Chaney, B. (2017). *Value-added evaluation of teacher preparation programs: Sensitivity of rankings to model specification*. Center for the Study of Testing, Evaluation, and Education Policy- Boston College.

Braun, H. I., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309-2344.

Braun, H. I., & Wainer, H. (2007). Value-added modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 27, (pp. 867-892). Elsevier Science. https://doi.org/10.1080/09332480.2011.10739845

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*, 199–231. https://doi.org/10.1214/ss/1009213726

Carnegie, N., Harada, M., & Hill, J. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness, 9*(3), 395-420. https://doi.org/10.1080/19345747.2015.1078862

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. The Quarterly Journal of Economics, 126(4), 1593-1660. https://doi.org/10.1093/qje/qjr041

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, R. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8-15. https://doi.org/10.1177/003172171209300603

Dearing, D., & Zachrisson, H. D. (2019). Taking selection seriously in correlational studies of child development: A call for sensitivity analyses. *Child Development Perspectives*, *13*(4), 267-273. https://doi.org/10.1111/cdep.12343

Diaconu, D. V. (2012). *Modeling science achievement differences between single-sex and coeducational schools: analyses from Hong Kong, SAR and New Zealand from TIMSS 1995, 1999, and 2003*. [Unpublished doctoral dissertation]. Boston College.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734

Freedman, D. A. (2010). *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815874

Ginsberg, A. L., Noel, J., & Plisko, V. W. (1988). Lessons from the Wall Chart. *Education Evaluation and Policy Analysis, 10*(1), 1-12. https://doi.org/10.2307/1163860

Goldhaber, D., Holden, K. L., & Grout, C. (2019). Errors in administrative education data: A cautionary tale. *Educational Researcher, 48*(3), 179-182. https://doi.org/10.3102/0013189X19837598

Hargreaves, A., & Braun, H. I. (2012). *Leading for All: A research report of the development, design, implementation and impact of Ontario's "Essential for Some, Good for All" initiative*. Council of Ontario Directors of Education. https://doi.org/10.1108/JPCC-06-2019-0013

Harris, D. (2009). Would accountability based on teacher value-added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*(4), 319-350. https://doi.org/10.1162/edfp.2009.4.4.319

Hedges, L. (1988). The meta-analysis of test validity: Some new approaches. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 191-212). Lawrence Erlbaum Assoc.

Holland, P. W., & Wainer, H. (1990). Sources of uncertainty often ignored in adjusting state mean SAT scores for differential participation rates: The rules of the game. *Applied Measurement in Education, 3*(2), 167-184. https://doi.org/10.1207/s15324818ame0302_3

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 697-701. https://doi.org/10.1371/journal.pmed.0020124

Jankowski, N., & Marshall, D. W. (2017). *Degrees that matter: Moving higher education to a learning systems paradigm*. Stylus Publishing, LLC.

Kalbfleisch, J. D., & Lawless, J. F. (1989). Inference based on retrospective ascertainment: An analysis of data in transfusion related AIDS. *Journal of the American Statistical Association, 84*(406), 360-372. https://doi.org/10.2307/2289919

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). *Doing data science: A framework and case study*. Harvard Data Science Review. https://doi.org/10.1162/99608f92.2d83f7f5

Kennet, R. S., & Redman, T. C. (2019). *The real work of data science: Turning data into information, better decisions, and stronger organizations*. Wiley. https://doi.org/10.1002/9781119570790

Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. Harvard Data Science Review. https://doi.org/10.1162/99608f92.17405bb6

Lockwood, J., & McCaffrey, D. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1*, 223-252. https://doi.org/10.1214/07-EJS057

Maronna, R. A. (2018). *Robust statistics: Theory and methods* (2nd ed.). Wiley. https://doi.org/10.1002/9781119214656

Mayo, D. G., & Cox, D. R. (2010). Frequentist statistics as a theory of inductive inference. In D. G. Mayo & A. Spanos (Eds.), *Error and Inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. Cambridge University Press.

McCaffrey, D. F., & Culpepper, S. A. (2021). Introduction to JEBS special issue on NAEP linked aggregate scores. *Journal of Educational and Behavioral Statistics, 46*(2), 135-137. https://doi.org/10.3102%2F10769986211001480

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics, 12*(2), 685-726. https://doi.org/10.1214/18-AOAS1161SF

Montgomery, M. R., Richards, T., & Braun, H. I. (1986). Child health, breastfeeding and survival in Malaysia: A random effects logit approach. *Journal of the American Statistical Association, 81*(394), 297-309. https://doi.org/10.1080/01621459.1986.10478273

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2001). The long-term effects of small classes in early grades: Lasting benefits in mathematics achievement at grade nine. *Journal of Experimental Education, 69,* 245-257. https://doi.org/10.1080/00220970109599487

Organisation for Economic Co-operation and Development. (2013). *Survey of adult skills technical report* (2nd ed.). OECD Publishing.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics, 37*, 187–204. https://doi.org/10.1080/07350015.2016.1227711

Pearman II, P. A., Springer, M. P., Lipsey, M., Lachowicz, M., Swain, W., & Farran, D. (2020). Teachers, schools, and pre-K effect persistence: An examination of the sustaining environment hypothesis. *Journal of Research on Educational Effectiveness, 13*(4), 547-573. https://doi.org/10.1080/19345747.2020.1749740

Provasniak, S. (2021). Process data, the new frontier for assessment development: rich new soil or a quixotic quest? *Large-scale Assessments in Education, 9*(1). https://doi.org/10.1186/s40536-020-00092-z

Reardon, S. F., Kalogrides, D., & Ho, A. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics, 46*(2), 138-167. https://doi.org/10.3102/1076998619874089

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519. https://doi.org/10.1162/edfp.2009.4.4.492

Rosenbaum, P. R. (1989). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika, 74*, 13- 26. https://doi.org/10.1093/biomet/74.1.13

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/978-1-4757-3692-2

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved bivariate covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological), 45*(2), 212-218. https://doi.org/10.1111/j.2517-6161.1983.tb01242.x

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571. https://doi.org/10.1162/edfp.2009.4.4.537

Russell, M., Ludlow, L., & O'Dwyer, L. (2019). Preparing the next generation of educational measurement specialists: A call for programs with an integrated scope and sequence. *Educational Measurement: Issues and Practice, 38*(4), 78-86. https://doi.org/10.1111/emip.12285

Sanders, W., Saxton, A., & Horn, S. (1997). The Tennessee value-added assessment system: A quantitative, outcome-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Corwin Press. https://doi.org/10.3102/10769986029001037

Shavelson, R. J., & Towne, L. (Eds.) (2002). *Scientific research in education.* National Academies Press. https://doi.org/10.17226/10236

Singer, J. D., Braun, H. I., & Chudowski, N. (Eds.) (2018). *International education assessments: Cautions, conundrums, and common sense.* National Academy of Education. https://doi.org/10.31094/2018/1

Singer, J. D., & Willett, J. B. (1990). Improving the teaching of applied statistics: Putting the data back into data analysis. *The American Statistician, 44*(3), 223-230. https://doi.org/10.2307/2685342

Stark, P. B., & Saltelli, A. (2018, August). Cargo-cult statistics and scientific crisis. *Significance*, *15*(4), 40-43. https://doi.org/10.1111/j.1740-9713.2018.01174.x

Steelman, L. C., & Powell, B. (1985). Appraising the implications of the SAT for education policy. *Phi Delta Kappan, 67*, 603-606.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact postselection inference for sequential regression procedures. *Journal of the American Statistical Association, 111*(514), 600-620. https://doi.org/10.1080/01621459.2015.1108848

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics, 39*(6), 478-501. https://doi.org/10.3102/1076998614558486

Tipton, E., & Olsen, R. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher, 47*(8), 516-524. https://doi.org/10.3102/0013189X18781522

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67. https://doi.org/10.1214/aoms/1177704711

Tukey, J. W. (1976). *Exploratory data analysis*. Addison-Wesley.

van der Sluijs, J., Craye, M., Funtowicz, S., Kloprogge, P., Ravetz, J., & Risbey, J. (2005). Combining quantitative and qualitative measures of uncertainty in model based environmental assessment: *The NUSAP System. Risk Analysis, 25*(2), 481-492. https://doi.org/10.1111/j.1539-6924.2005.00604.x

von Davier, M., & Sinharay, S. (2014). Analytics in international large scale assessments: item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155-174). CRC Press.

Wainer, H. (1986a). The SAT as a social indicator. A pretty bad idea. In H. Wainer (Ed.) *Drawing Inferences from self-selected samples* (pp. 7-21). Springer-Verlag. https://doi.org/10.1007/978-1-4612-4976-4_2

Wainer, H. (1986b). Five pitfalls encountered when trying to compare states on their SAT scores. *Journal of Educational Statistics, 11*, 239-244. https://doi.org/10.1111/j.1745-3984.1986.tb00235.x

Walsh, M. E., Madaus, G. F., Raczek, A. E., Dearing, E., Foley, C., An, C., Lee-St. John, T. L., & Beaton, A. E. (2014). A new model for student support in high-poverty urban elementary schools: Effects on elementary and middle school academic outcomes. *American Educational Research Journal, 51*(4), 704-737. https://doi.org/10.3102/0002831214541669

Wolf, R., Morrison, J., Inns, A., Slavin, R., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness, 13*(2), 428-447. https://doi.org/10.1080/19345747.2020.1726537