Special Issue Article

Ethnography

Ethnography
2018, Vol. 19(2) 227–253

DOI: 10.1177/1466138117725341
journals.sagepub.com/home/eth
SAGE

Aligning sampling and case selection in quantitative-qualitative research designs: Establishing generalizability limits in mixed-method studies

Bryan L Sykes
University of California Irvine, USA

Anjuli Verma
University of California Berkeley, USA

Black Hawk Hancock
DePaul University, USA

Abstract
Quantitative researchers increasingly draw on ethnographic research that may not be generalizable to inform and interpret results from statistical analyses; at the same time, while generalizability is not always an ethnographic research goal, the integration of quantitative data by ethnographic researchers to buttress findings on processes and mechanisms has also become common. Despite the burgeoning use of dual designs in research, there has been little empirical assessment of whether the themes, narratives, and ideal types derived from qualitative fieldwork are broadly generalizable in a manner consistent with estimates obtained from quantitative analyses. We draw on simulated and real-world data to assess the bias associated with failing to align samples across qualitative and quantitative methodologies. Our findings demonstrate that significant bias exists in mixed-methods studies when sampling is incongruent within research designs. We propose three solutions to limit bias in mixed-methods research.

Keywords
mixed methods, generalizability, validity, replication, sampling, qualitative, quantitative, bias, triangulation

Corresponding author:
Bryan L. Sykes, University of California, 3317 Social Ecology II, Irvine, CA 92697, USA.
Email: blsykes@uci.edu

Introduction

Simmering debates among ethnographers about representation, reproducibility, generalizability and the role of theory have come to a boil over the past decade. Scholars who study social inequality and mass incarceration from a variety of methodological orientations confront related tensions around validity, replication and generalizability in research that blends quantitative and ethnographic methods (see e.g. Cohen, 2015; Duneier, 2004, 2006; Goffman, 2014; Katz, 1997; Klinenberg, 2002, 2004, 2006; Platt, 2016; Ralph, 2015; Rios, 2015; Sánchez-Jankowski, 2002; Sharkey, 2015; Venkatesh, 2013; Wacquant, 2002; Wilson, 2014). As quantitative-qualitative research designs become increasingly common and seen as value added for policy-relevant studies of race, law, and inequality, the pot may have boiled over. Even as these debates continue to spill beyond the pages of scholarly journals and into the popular press and public domain (e.g. Lubet, 2015; Cohen, 2015; Lewis-Kraus, 2016; Platt, 2016), a core set of questions about the possibilities and limits of translation, integration, and commensurability across methods and disciplines, as well as *within* research designs, remains decidedly unsettled in the social science domain, perhaps even more so among ethnographers and quantitative researchers with explicit commitments to methodological 'pluralism' and the constructive value of working across disciplinary divides (Guba and Lincoln, 2005; Lamont and Swidler, 2014: 153; Lamont and White, 2009).

This article takes on a timely and important question: Can mixed-methods studies deliver on their promise to provide a fully-rounded view of the social world? Our particular aim here is to examine the possibility that researchers will be misled by mixing samples derived from probabilistic and non-probabilistic sampling. Does combining non-probabilistic qualitative case data with probabilistic survey data affect inferences in mixed-methods studies, and if so, how biased are results?

The pragmatism that animates a growing and increasingly formalized field of mixed-methods research (henceforth referred to as 'MMR') has also infused ethnographic social inequality research (e.g. Small, 2013; Tavory and Timmermans, 2013), where interdisciplinarity and innovation are increasingly favored over methodological 'tribalism' (Lamont and Swidler, 2014: 153). Researchers may be trained to understand science as a 'conversation' (Abbott, 2004: 3), in which 'translation' across multiple 'languages' (Small, 2009: 10; see also Blalock, 1968; Small, 2011) and 'mutual methodological critique' (Abbott, 2004: 75) rather than competition produces knowledge. Additionally, the sociologist-as-public intellectual and attendant pull of the policy audience in decisions about how to 'pragmatically' design, conduct, and communicate findings from research on systems of inequality (see e.g. Burawoy, 2005) also seems to be shaping a sensibility among some researchers that two methods are better than one and that mixed-methods studies enhance the validity of findings (Greene, 2007; O'Cathain and Thomas, 2006; Pearce, 2002).

It is, therefore, not surprising that recent social inequality research has taken a mixed-methods turn, in which quantitative researchers may draw on ethnographic accounts to add depth to their analyses of large datasets, and ethnographers may

use quantitative data to add breadth to their findings in a given case. Quantitative researchers increasingly draw on ethnographic research that may not be generalizable to inform and interpret results from statistical analyses; at the same time, while generalizability is not always an explicit ethnographic research goal, the integration of quantitative data by ethnographic researchers to buttress findings on processes and mechanisms has also become increasingly common (Leahey, 2007; Pearce, 2002; Small, 2009, 2011; Small et al., 2008). In the social inequality field, for example, many ethnographic studies have come to use nationally (or locally) representative quantitative data as a way of addressing limitations in the generalizability of case-specific findings, as well as to bolster the broader relevance and validity of research as interpreted in the public imagination and policy field. However, despite the burgeoning use of dual research designs, there has been little empirical assessment of whether the themes, narratives, and ideal types derived from qualitative fieldwork are broadly generalizable in a manner consistent with estimates obtained from quantitative analyses based on sample surveys.

Against the backdrop of recent (and recurring) controversies concerning race, law, and social inequality research that blends quantitative and qualitative methods (e.g. Goffman, 2014; Klinenberg, 2002), we offer a systematic examination of the conditions under which qualitative-quantitative research designs undermine scientific definitions of validity. We draw attention to a fundamental methodological tension that is often overlooked in these debates, which sets the stage for subsequent, overarching disputes and dilemmas of translation, integration, and commensurability: *the alignment of sampling and case selection procedures in mixed-methods research designs*. We argue that 'alignment' occurs when the underlying population distribution is the same for samples used in qualitative and quantitative analyses within a mixed-methods study. We demonstrate that bias exists within some forms of 'concurrent' sampling when quantitative and qualitative data are 'mixed' without adjusting the case sampling weights for differential probabilities of selection in dual design research. Findings from our analyses have implications for theory construction, hypothesis testing, and the generalizability of mixed-methods. As a result, we offer three types of sampling solutions and adjustments to sampling weights for researchers interested in conducting mixed-method studies: *prospective sampling, retrospective sampling*, and *aligned concurrent sampling*.

Sampling in ethnographic mixed-methods research

Sampling and case selection procedures are first-order research design decisions that chart the course for the depth and/or breadth of future findings. 'In research, sampling is destiny' (Kemper et al., 2003: 275), much like 'the selection of the basic object of analysis' (Desmond, 2014: 547). In any study, sampling is the primary stage in measuring this object of analysis. The sampling unit must first be specified and then the decision about how to select which of those units and how many to

include become foundational premises of the ethnography. In selecting units, one may make conscious decisions about which units *not* to study, as well as unconscious omissions of units that may exist in a universe that the ethnographer cannot fully conceive at the start of the study.

The dichotomy between probabilistic and non-probabilistic sampling procedures used in quantitative and ethnographic research, respectively, reflects the distinct epistemological and ontological premises of each methodological orientation. However, the mixed-methods turn has begun to subvert this dichotomy through an emergent species of studies that blend probabilistic and non-probabilistic sampling schemes in problematic ways (Tashakkori and Teddlie, 2003). Teddlie and Yu (2007: 85) conceptualize a 'Purposive-Mixed-Probability Sampling Continuum' based on the extent to which sampling procedures are integrated across the quantitative (QUAN) and qualitative (QUAL) strands of the research design. Notwithstanding the 'lack of details regarding sampling in many' (p. 91) of the numerous MMR articles they surveyed, Teddlie and Yu (2007: 85) describe the general approach to sampling in MMR:

> The MM [mixed-methods] researcher sometimes chooses procedures that focus on generating representative samples, especially when addressing a QUAN strand of a study. On the other hand, when addressing a QUAL strand of a study, the MM researcher typically utilizes sampling techniques that yield information rich cases. Combining the two orientations allows the MM researcher to generate complementary databases that include information that has both depth and breadth regarding the phenomenon under study.

Teddlie and Yu (2007: 89) then offer a provisional classification of MMR sampling strategies. Most common is the 'sequential' technique, where 'information from the first sample (typically derived from a probability sampling procedure) is required to draw the second sample (typically derived from a purposive sampling procedure) (Kemper et al., 2003; Leahey, 2007). For example, in Pearce's (2002: 104) study of the influence of religion on childbearing preferences, survey analysis and sampling techniques were used to systematically select 'anomalous' cases for ethnographic study. This mixed-method, deviant case analysis ultimately led Pearce to revise theories, code new survey measures, and improve the fit of statistical models.

Contrary to Pearce's design, 'concurrent' sampling 'involves the selection of units of analysis for an MM study through the simultaneous use of both probability and purposive sampling. One type of sampling procedure does not set the stage for the other … instead, both probability and purposive sampling procedures are used at the same time' (Teddlie and Yu, 2007: 89). In other words, sampling in the QUAN and QUAL strands of the study occur independently. Such 'concurrent' designs purportedly enable researchers to 'triangulate' results from different strands of the study to 'confirm, cross-validate, or corroborate findings within a single study' (Creswell et al., 2003: 299).

It is this 'concurrent' form of MMR – where sampling in QUAN and QUAL research strands remain independent yet serve as the basis for comparison – with which we take issue and focus our critique. The notion that quantitative and qualitative research can be 'mixed' or used to 'triangulate' when findings are derived from separate samples selected according to incommensurate sampling procedures (either probabilistic or non-probabilistic) is mistaken.

We contend that if the goal of an MMR study is generalizability and consistency, the sampling strategies used in qualitative and quantitative research strands must be aligned – that is, samples should be drawn from the same known population distribution. Failure to align the samples according to the same population distribution will produce inconsistent and biased estimates in these kinds of MMR designs, potentially undercutting findings from both methods. For MMR to be a valid hybrid that optimizes both qualitative and quantitative epistemologies, qualitative research must be reproducible and employ probability-based sampling techniques (Lucas, 2014), and quantitative researchers must consider whether unweighted observations from qualitative research they wish to integrate should be adjusted based on the sampling weights in national or local data. Failure to align the underlying population distribution from which samples are drawn in qualitative and quantitative research designs may alter the correlational importance of some processes and mechanisms that were uncovered from field observations. In quantitative research, hypothesis testing serves as the basis for theory and mechanistic evaluations. Mechanisms and processes may be unobservable in quantitative data but observable in qualitative data (e.g. Pearce, 2002). When quantitative researchers wish to test hypotheses concerning mechanisms and processes derived from qualitative research, they risk reproducing unobserved bias associated with non-probability based, purposive or convenience sampling.

A theoretical demonstration of the problem

We conceptualize the 'synergistic' (Hall and Howard, 2008) promise of MMR to deliver a fully-rounded view of the social world as *breadth-depth optimization*: a dual research design that integrates the breadth gained from quantitative data with the depth uniquely gained from ethnographic and qualitative data. Dual designs that maximize the number of randomly-selected cases (breadth) and the quality of information derived from cases (depth) represent the optimal point of methodological integration. Quantitative data (e.g. survey data) often uses large-n datasets based on randomly-sampled units so that causal inferences and associations can be drawn and generalized to a larger population; qualitative lines of inquiry (e.g. ethnographies) proceed from constructivist, interpretivist or hermeneutic paradigms, wherein the existence of an objective, knowable, or fixed 'truth' is questioned and viewed as dependent on situated or subjective interpretations (Haraway, 1988; Small, 2011).

The goal of MMR is to maximize 'both depth and breadth regarding the phenomenon under study' (Teddlie and Yu, 2007: 85). Figure 1 depicts the relationship

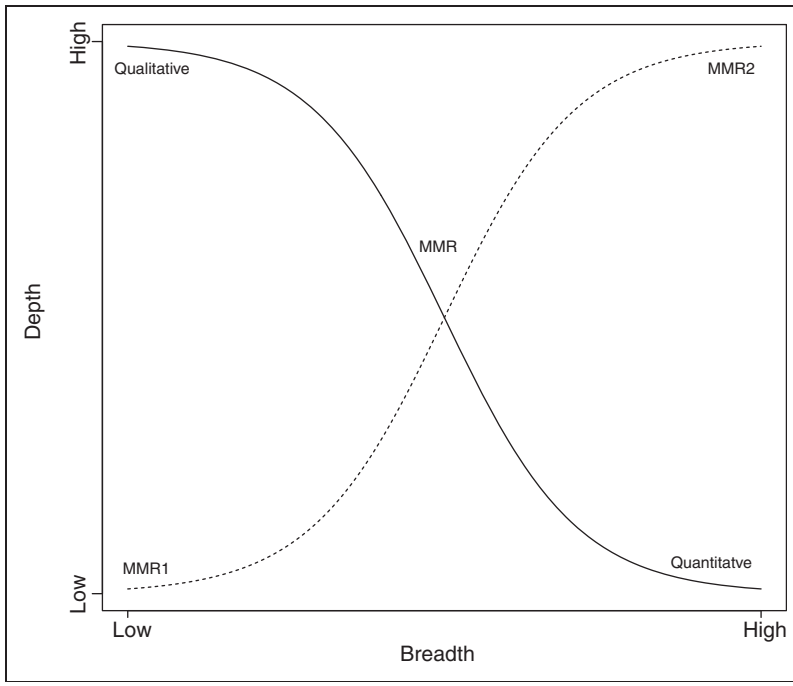

Figure 1. Breadth-depth optimization in mixed-methods research designs.

between breadth (x-axis) and depth (y-axis) as research objectives in mono-method (purely qualitative or quantitative) and mixed-method research designs. The solid line represents conventional beliefs about the necessary quantitative-qualitative tradeoff between breadth and depth in mono-method research: breadth and depth are inversely related, where a gain in one dimension corresponds to a non-linear loss in the other. In this framework, the optimal point for integrating qualitative and quantitative designs, labeled 'MMR,' is balanced between the low and high points of the breadth-depth continuum.

Yet, we contend that the breadth-depth tradeoff function of fully integrated mixed-methods designs, depicted by the dashed line, is the chiral (or mirror) image of mono-method designs. The dashed line highlights the promises and perils of mixed-method designs due to sampling procedures. When the underlying population distribution is the same for sampling procedures used across qualitative and quantitative methods, the breadth-depth optimization potential of research designs can transcend the 'MMR' mid-point and reach the maximum optimization point, labeled 'MMR2' on the dashed line. At this point, mixed-methods research can deliver on the promise to provide a fully-rounded view of the social world. However, when unaligned sampling in mixed-methods research leads to incorrect conclusions that have implications for theory generation, hypothesis testing, and generalizability, the value added from attempting to maximize both breadth and

depth may lead to lower quality research that provides a perilously partial view of the social world despite multiple methods, dragging the optimization potential to its lowest point, labeled 'MMR1' on the dashed line.

Taken together, these non-linear curves allow for a comparison of the breadth-depth optimization across research designs, revealing that only a subset of mixed-methods designs can conceptually generate the presumed value added of integrating quantitative and qualitative methodologies: only the 'MMR2' designs approach maximum breadth *and* depth. By contrast, the 'MMR' designs appear to offer no more optimization over the use of either qualitative or quantitative mono-method designs, as they reach the limit of breadth-depth optimization at the same point. The 'MMR1' mixed-methods designs fall short of both purely qualitative *and* purely quantitative designs because they can attain only minimal breadth and depth. As we will show in subsequent analyses, researchers can reach erroneous conclusions when their findings are based on samples selected according to different concepts of the underlying population distribution across qualitative and quantitative methods.

Pearce's (2002) study exemplifies what we articulate as 'MMR2' in Figure 1. Because the quantitative and qualitative samples in her study were drawn from the same population distribution, she was able to identify and contextualize ethnographic findings from anomalous cases. The value of the depth she gained from qualitatively examining cases allowed her to refine theoretical explanations for the phenomenon under study, as well as to improve existing statistical models. Had she failed to align sampling across methods, the insights she stood to gain through the ethnographic case studies would be limited to those cases alone ('MMR'). Worse, if she tried to relate these ethnographic findings to her quantitative findings without knowing whether her cases deviated from or represented the norm, she would have risked arriving at erroneous conclusions, which we argue is suboptimal ('MMR1') to simply conducting a mono-method study.

An empirical demonstration of the problem

Data and methods

Parental incarceration is used as a heuristic to demonstrate how findings from a hypothetical mixed-methods study on mass incarceration and social inequality can reach erroneous conclusions when sampling procedures are not consistent. Data from the 2011–12 National Survey of Children's Health (NSCH) are employed to investigate racial differences in parental incarceration. These data are collected for the Centers for Disease Control (CDC) by the National Opinion Research Center (NORC) at the University of Chicago. The NSCH randomly samples telephone numbers to locate households with children aged 0–17, and within each household one child was randomly selected to be the subject of interview.[1] These data have been used by mass incarceration researchers to investigate the risk of exposure to parental incarceration (Sykes and Pettit, 2014), social program enlistment due to

parental incarceration (Sykes and Pettit, 2015), and the association between parental incarceration and child health (Turney, 2014).

The NSCH asked parents a variety of questions about family functioning, parental health, and social background characteristics. Interviewers completed 95,677 child-level interviews, with the number of interviews ranging from over 1800 to 2200 per state. Weighted (or aligned) estimates represent the social experiences of all non-institutionalized minors in the United States. For expositional purposes, we retain only race and parental incarceration status from these data to draw attention to how sampling in non-probabilistic fieldwork and interviews has significant import for evaluating mechanisms and processes in MMR – for both quantitative and qualitative researchers.

We supplement the NSCH data with synthetic data from a Monte Carlo simulation (see Diaconis, 2009; Lucas and Szatrowski, 2014). Monte Carlo (MC) methods represent a class of algorithms that allow the analyst to draw samples from a sampling distribution, enabling the analyst to 'summarize the theoretical distribution using these simulated values' (Gill, 2002: 239). The use of MC methods is important because we can: '1) specify a true model; 2) generate a dataset using this true model; and 3) calculate the test statistic or estimator that is being evaluated with this artificially generated sample and store the results' (Johnston and Dinardo, 1997: 348). First, we use MC methods to simulate the process of 'saturation' by setting a hypothetical number of interviews or field observations qualitative researchers may require to achieve stability in their findings with respect to coded themes, ideal types, and narratives. We define case saturation as the number of observations made by the time researchers exit the field, or when researchers believe that they have achieved a saturated knowledge about their case and its narratives. Second, we code those narratives across each respondent-interview (or respondent-observation) in order to make meaning of the themes that emerge to highlight or underscore the processes and mechanisms that explain social inequality (e.g. class background, neighborhood attributes, and institutional relationships).

We begin by drawing a sample size of 95,677 (one for each child in the NSCH) from a Poisson distribution. In the field, qualitative researchers do not know the distributional properties underlying the number of observations that comprise case saturation. It could be, for instance, that the range (or variance) of observations necessary to reach saturation exceeds the mean number of observations, resulting in overdispersion (Long, 1997), which confounds research that relies on counting methods. This overdispersion must be corrected before drawing inferences in both qualitative and quantitative research. The Poisson distribution has a number of useful features that we exploit. First, for the purposes of saturation, qualitative researchers are generally concerned with how many cases they may need, how many interviews to conduct, or how many times to observe an environment. Because narrative saturation varies across cases and field sites, coded themes and ideal types have a random distribution of their own that has been largely ignored in the discussions of case saturation. If the number of observations required for narrative saturation in any case is (y), then (y) is a random variable with a mean count

(μ) greater than 0, such that the number of observations during an interval of time can be sampled from a Poisson distribution, where

$$\Pr(y|\mu) = \frac{\exp(-\mu)\mu^{y}}{y!} \quad \text{for } y = 0, 1, 2, \ldots$$

As the mean count (μ) increases, the Poisson distribution shifts to the right (Long, 1997), such that the mean and variance are *equidispersed*, ensuring that we do not need to correct for overdispersion in mixed-methods models, as displayed by

$$\mathrm{Var}(y) = \mathrm{E}(y) = \mu$$

Additionally, as the mean count of observations (μ) increases, the probability of having 0's (unobserved or non-response cases) decreases and approximates the normal distribution. We set μ equal to 4 for our simulated sample of observations as a hypothetical number of field observations required for saturation. The number of possible observations or interviews ranges from 0 to 15. Past ethnographic research has observed field sites as few as five times (Hancock, 2013). Mixed-methods studies have planned an average of four follow-up interviews (Western et al., 2014), and another mixed-methods study made field observations as many as eight times for some cases (Small et al., 2008). Thus, the range and mean for our simulated model is realistic based on some notable studies. Figure 2 displays the number and distribution of observations from our simulation.

A second benefit of the Poisson distribution is that we can easily compare theoretical and observed sample properties. Evans, Hastings and Peacock (2000: 155) argue that because of equidispersion, the mean and variance can be estimated by 'observing the characteristics of actual samples.' A comparison of theoretical and actual sample statistics will allow us to estimate the bias associated with not calibrating sampling differences between qualitative-quantitative research designs.

Finally, for each respondent-observation (or respondent-interview), we randomly assign a thematic narrative to that in-depth field observation (or interview).[2] We set the hypothetical number of possible thematic narratives to five: A, B, C, D, and E. Each theme has an equal probability of assignment.[3]

In summary, our steps in constructing the mixed-methods dataset proceeded as follows. First, there are 95,677 observations in the NSCH. We drew a sample of 95,677 numbers – one for each respondent in the NSCH – from the Poisson distribution to simulate ethnographic field observations and case saturation.[4] We set the average number of field observations at 4, which means that each respondent is observed 0 to 15 times. Respondents who were randomly assigned a 0 for their field observation represent the unobserved or non-respondent case.[5] Second, among the respondents who were observed, each observation was coded as 1 of 5 themes: A, B, C, D, or E, with each theme having an equal probability of assignment. Finally, we coded the most prevalent theme among respondent-observations. We used sample weights from the NSCH to estimate population descriptive statistics and

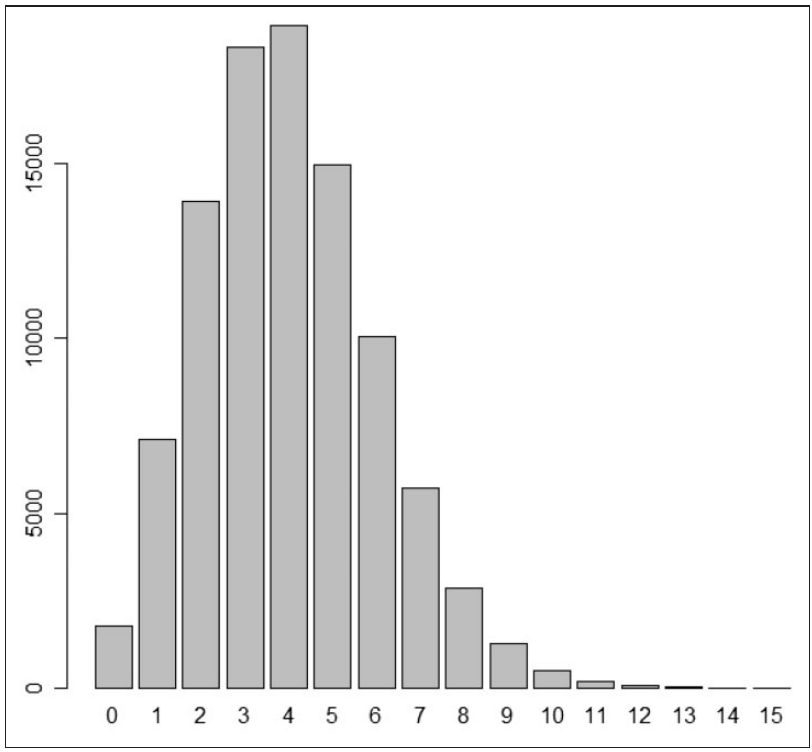

Figure 2. The synthetic number of interviews (or field observations) necessary for thematic narrative saturation.
Source: Authors' calculations of a simulated dataset from a Poisson distribution with a mean of 4. The seed was set at 1001 (in R).

to estimate causal relationships, which we argue is the true representation of the social world under examination.[6]

Hypotheses

Because the probability of selecting a particular case in the field may be unknown, comparisons of narratives between cases are usually given equal weight in their representation within the population (see Lucas, 2014: 395 for discussion on equal weights). When combined with quantitative samples drawn from a population, the sampling weights between these two data sources may be unaligned. However, when quantitative and qualitative cases are sampled in the same way, the sampling weights will be aligned. We use weights from NSCH data to estimate true sample statistics for mixed-methods findings. Thus, if qualitative and quantitative data are sampled in similar ways, we would expect our narratives to be 1) consistently positive or negative when comparing unaligned and aligned

coefficients and 2) consistently statistically significant or insignificant between unaligned and aligned models.

Measures

Table 1 displays the operationalization and population descriptive statistics in our study. Non-Hispanic white children represent 53% of respondents, whereas non-Hispanic black, Hispanics, and non-Hispanic other children account for 14%, 24%, and 10%, respectively.

Respondents synthetically completed an average (μ) four in-depth interviews or observations. Although each thematic narrative associated with a respondent-observation had an equal chance of selection, the prevalence of a particular theme within the population is not uniform. Of the thematic narratives that represent potential processes and mechanisms for parental incarceration, 36% of respondent narratives fall into Theme A, whereas themes B, C, D, and E respectively account for 21%, 17%, 13%, and 13% of the possible thematic codings.

Findings

Table 2 presents estimates from an OLS model predicting the number of interviews (or observations) from aligned and unaligned models. Because we randomly sampled the number of interviews (or observations) from the Poisson distribution and assigned them to each of the respondents in the NSCH data, statistics from unaligned models highlight the theoretical associations of social background characteristics on the number of interviews, while statistics from aligned models represent the 'actual' associations of social background in the real world. There are three important findings from this table. First, our simulated cases are indeed random, as unaligned estimates are not statistically associated with the outcome, and independent variables explain virtually none of the variation in the number of interviews or observations.

The second finding is that we replicate the general conditions of statistical bias associated with non-probability based sampling identified in the literature. Using a different estimation strategy, Lucas (2014: 399) shows that non-probability based samples are 'horribly biased' because 'high and low estimates never bracket the population value.' We find that two of seven statistical estimates reverse in direction. The positive association between the number of interviews, Hispanic background, and having experienced parental incarceration changes to negative when probability sampling weights are implemented.

The third finding is that the level of significance also changes in relation to probability-based sampling. Although Lucas (2014) shows how the correlation coefficients change under repeated sampling, we show that the problem extends to hypothesis testing. Two of seven independent variables (Hispanic and the interaction effect between Hispanic and parental incarceration) are statistically significant when official sampling weights are used to adjust for national representation.[7]

Table 1. Descriptive statistics of variables and operationalizations, NSCH 2011–2012 & simulated data.

Variables	Operationalization	Coding	Mean	SD	Min	Max
Parental Incarceration	Whether the child has a parent who has been incarcerated	Y = 1, N = 0	0.07	0.25	0	1
NH-White	Child is non-Hispanic white (baseline)	Y = 1, N = 0	0.53	0.50	0	1
NH-Black	Child is non-Hispanic black	Y = 1, N = 0	0.14	0.34	0	1
Hispanic	Child is Hispanic	Y = 1, N = 0	0.24	0.42	0	1
NH-Other	Child is non-Hispanic other	Y = 1, N = 0	0.10	0.30	0	1
Working Poor	Household is below the poverty line even though someone is employed full-time	Y = 1, N = 0	0.13	0.30	0	1
Observations/Interviews	Number of field observations or in-depth interviews for the i-th respondent (drawn from Monte Carlo simulation of Poisson Distribution)	# of obs.	4.00	2.00	0	15
Thematic-Narrative: A	Recode of Most Prevalent Theme #1 for i-th respondent, Randomly Drawn with Equal Chance of Selection (1 of 5)	Y = 1, N = 0	0.36	0.48	0	1
Thematic-Narrative: B	Recode of Most Prevalent Theme #2 for i-th respondent, Randomly Drawn with Equal Chance of Selection (2 of 5)	Y = 1, N = 0	0.21	0.40	0	1
Thematic-Narrative: C	Recode of Most Prevalent Theme #3 for i-th respondent, Randomly Drawn with Equal Chance of Selection (3 of 5)	Y = 1, N = 0	0.17	0.37	0	1
Thematic-Narrative: D	Recode of Most Prevalent Theme #4 for i-th respondent, Randomly Drawn with Equal Chance of Selection (4 of 5)	Y = 1, N = 0	0.13	0.33	0	1
Thematic-Narrative: E	Recode of Most Prevalent Theme #5 for i-th respondent, Randomly Drawn with Equal Chance of Selection (5 of 5)	Y = 1, N = 0	0.13	0.33	0	1

Source: Authors' calculation of National Survey of Children Health (NSCH) and simulated data. NSCH-weighted N = 73,716,871

Table 2. Estimates from an OLS model predicting the number of field observations necessary for narrative saturation.

	Number of Interviews (or Field Observations)	
	Unaligned	*Aligned*
NH-Black	0.026	0.027
	(0.024)	(0.048)
Hispanic	0.008	−0.093*
	(.020)	(.043)
NH-Other	0.015	0.050
	(0.022)	(0.050)
Parental Incarceration	0.024	−0.043
	(0.036)	(0.071)
NH-Black x Parental Incarceration	0.071	0.187
	(0.079)	(0.140)
Hispanic x Parental Incarceration	0.034	0.417*
	(0.078)	(0.170)
NH-Other x Parental Incarceration	0.112	0.097
	(0.080)	(0.150)
R^2	0.01%	0.10%
Number of Directionality Changes	2/7	
Number of Significance Changes	2/7	

^p < .10, *p < . 05, **p < . 01, ***p < .001
Source: Authors' calculation from National Survey of Children's Health (NSCH) and Monte Carlo simulated data.
Note: Non-Hispanic whites are the reference groups.

We explore the bias associated with failing to align the sampling distributions in several ways. First, we investigate how treating each thematic narrative independent from a true sampling population with known weights affects the overall pattern of bias in explaining parental incarceration. The bias inherent in these themes grossly distorts our proposed theoretical mechanism. Figure 3 depicts the total bias associated with not adjusting for population sampling. Thematic narrative A contains 10.8% bias, while thematic narratives B, C, D are biased by 13.9, 13.0, 16.5%, respectively.

To understand fully the sources of bias in these statistics, Table 3 disaggregates the narrative distortions in rates of parental incarceration by race. Because the total bias for each theme is an aligned average of the racial distribution (from Table 1) and sampling bias (from Figure 3), we can disaggregate the total distortion into its

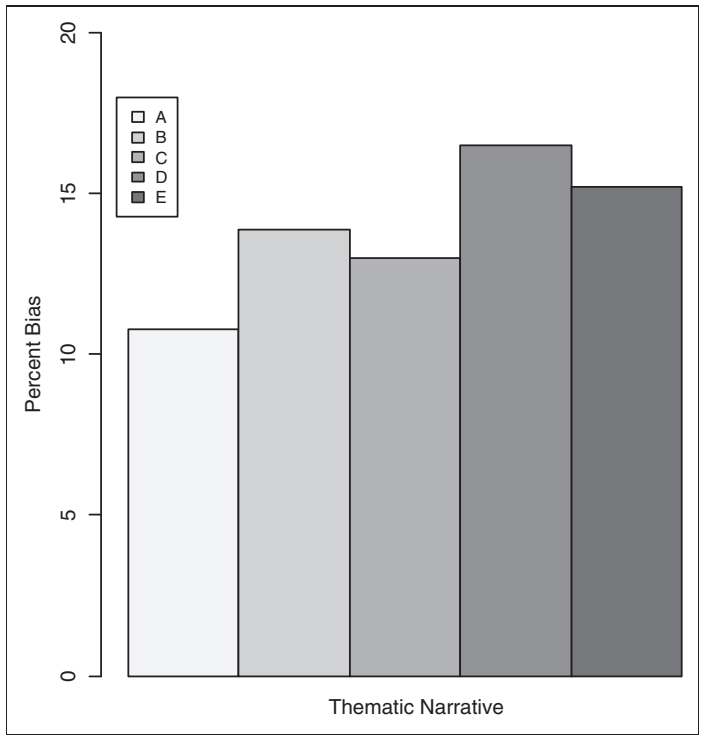

Figure 3. Percent bias from failing to align sampling distributions in thematic narratives that explain parental incarceration.
Source: Authors' calculation from National Survey of Children's Health (NSCH) and Monte Carlo simulated data.

Table 3. Percent bias from failing to align sampling distributions in thematic narratives that explain parental incarceration, by race.

	Thematic Narrative Bias					
	A	B	C	D	E	Number of Directional Changes
NH-White	17.3^{+}	17.5^{+}	17.7^{+}	15.7^{+}	16.8^{+}	5/5
NH-Black	19.8^{+}	13.5^{+}	6.6^{+}	19.8^{+}	11.2^{+}	5/5
Hispanic	29.1^{+}	2.0^{+}	5.9	3.2^{+}	10.2^{+}	4/5
NH-Other	11.1	11.1	19.1	0.3	35.8	0/5
Number of Directionality Changes	3/4	3/4	2/4	3/4	3/4	14/20

$^{+}$reversal in directionality of estimates
Source: Authors' calculation from National Survey of Children's Health (NSCH) and Monte Carlo simulated data.

race-specific components across themes. The estimate of parental incarceration associated with theme E is largest for non-Hispanic others; over one-third of this statistic is biased due to sampling. Non-Hispanic others also experienced the lowest level of bias (0.3%) for Theme D.

We also show the change in directionality across themes. Whites and blacks experienced the greatest number of directional changes (5 out of 5) of any racial group, and non-Hispanic others experienced the fewest (0 out of 5). Overall, 70% of cell entries exhibit some pattern of directionality reversal.

Similarly, thematic narratives also display a great deal of variability in directional changes across racial groups. The variability in estimate reversals of parental incarceration across racial groups varies from a low of 50% (Theme C for whites and blacks) to a high of 75% (Themes A, B, D, and E for whites, blacks, and Hispanics).

We also consider the practical consequences of this bias for narrative-specific outcomes. Table 4 estimates a multinomial logistic regression comparing the aligned and unaligned associations of race and parental incarceration between specific themes.[8] Specifically, we compare Theme E to all previous thematic narratives. We report estimates as relative risk ratios.

In comparing unaligned and aligned estimates from the multinomial model of Theme A vs. Theme E (i.e., A | E), we see that the number of relative risk ratios changes direction two out of seven times (from greater than 1 to less than 1, as well as from less than 1 to greater than 1). The largest difference is the comparison between B | E. Not only does the directionality of coefficients change two out of seven times, but the level of significance changes twice; two independent variables become insignificant (non-Hispanic black and its interaction with parental incarceration) once the probabilistic sampling weights are aligned. Directional comparisons for models C | E and D | E are also incorrect two out of seven times. Interestingly, three out of four estimates for Hispanics display issues with either statistical significance or directional associations (A | E, C | E, and D | E). Shockingly, the overall direction for the interaction between Hispanic and parental incarceration is wrong across all four models that rely on thematic narratives.

Finally, we analyze how the 'zeroth' or unobserved case matters for the assessment of social processes related to racial inequality in parental incarceration. Although many qualitative researchers rarely specify how many field observations were necessary to reach case saturation after the fact, or retrospectively, ethnographers increasingly draw attention to the importance of the zeroth, negative, or disconfirming case (Desmond, 2014; Small, 2009; Tavory and Timmermans, 2013). The zeroth, negative, or disconfirming case is the result of either non-response among subjects critical to the ethnography or from failure to reach narrative saturation from field observations. Because the mean number of observations necessary for narrative saturation in our models is drawn from a Poisson distribution – where some cases have zeros due to either non-response or failure to reach narrative saturation – we can simultaneously investigate how both non-random case selection and failure to explore unobserved and disconfirming cases matters for understanding mechanisms of social inequality. If these disconfirming

Table 4. Relative risk ratios from a multinomial logit model comparing the effect of race and parental incarceration on thematic narrative contrasts.

	A compared to E		B compared to E		C compared to E		D compared to E	
	Unaligned	*Aligned*	*Unaligned*	*Aligned*	*Unaligned*	*Aligned*	*Unaligned*	*Aligned*
NH-Black	0.98	0.92	**1.09***	**1.10**	0.99	0.97	1.00	0.99
	(0.04)	(0.07)	(0.05)	(0.08)	(0.04)	(0.08)	(0.05)	(0.09)
Hispanic	**1.05**	**1.14^**	1.05	1.09	1.01	0.99	**1.12****	**1.05**
	(0.04)	(0.08)	(0.04)	(0.09)	(0.04)	(0.08)	(0.04)	(0.09)
NH-Other	1.02	1.05	1.06	1.11	1.05	1.06	1.03	1.13
	(0.04)	(0.09)	(0.07)	(0.10)	(0.04)	(0.10)	(0.04)	(0.11)
Parental Incarceration	1.06	1.07	1.07	1.07	1.02	1.04	0.96	0.95
	(0.06)	(0.12)	(0.07)	(0.14)	(0.07)	(0.13)	(0.07)	(0.14)
NH-Black x Parental Incarceration	0.84	0.93	**0.69****	**0.70**	0.92	0.86	0.83	0.94
	(0.10)	(0.21)	(0.10)	(0.18)	(0.13)	(0.21)	(0.13)	(0.26)
Hispanic x Parental Incarceration	1.04	0.71	1.00	0.90	1.16	0.97	1.06	0.99
	(0.14)	(0.22)	(0.14)	(0.32)	(0.17)	(0.33)	(0.17)	(0.36)
NH-Other x Parental Incarceration	0.93	1.14	0.92	1.14	0.86	0.98	1.14	1.60
	(0.12)	(0.30)	(0.13)	(0.32)	(0.13)	(0.28)	(0.18)	(0.80)
Number of Directionality Changes (underlined)	2/7		2/7		2/7		2/7	
Number of Significance Changes (bolded)	1/7		2/7		0/7		1/7	

^p < .10 *p < . 05 **p < . 01 ***p < .001

Source: Authors' calculation from National Survey of Children's Health (NSCH) and Monte Carlo simulated data.

Note: Non-Hispanic whites and Thematic Narrative E are the baseline groups.

cases do not matter for drawing valid mixed-methods conclusions, we expect there to be no directional changes in the respondents' characteristics when the sampling distributions are aligned, and their levels of statistical association should not lose or gain significance after alignment.

Table 5 presents odds ratios from a logistic regression model predicting the zeroth or negative case. We also assess how household employment and poverty matter for the odds of a respondent being unobserved as a potentially disconfirming case. In models that draw on all cases, regardless of poverty and employment status among the primary parent/guardian, we find that non-random case selection does in fact result in directional changes of relationships between race, parental incarceration, and non-response in three out of seven instances. Specifically, non-Hispanic blacks and Hispanics (as compared to non-Hispanic whites) appear more likely to be overlooked as negative cases when sampling distributions are aligned. This pattern holds when the data are disaggregated by non-working poor households. Yet, among the working poor, the selection bias associated with unaligned sampling distributions produces two effects: first, the number and type of directionality changes vary across different racial categories (e.g. non-Hispanic blacks), and second, parental incarceration goes from being non-significant to significant once the sampling distributions are aligned between simulated and real data. Particularly noteworthy is that all models predicting the odds of being the zeroth case change in direction for non-Hispanic black children. Findings from this analysis highlight the importance of both selecting disconfirming cases and aligning the sampling distributions of qualitative and quantitative data to prevent drawing erroneous conclusions.

Three proposed solutions

So what can be done? For MMR studies with the goal of generalizability and maximum breadth-depth optimization (i.e. MMR2 in Figure 1), we propose three sampling procedures as possible solutions. The first is *prospective sampling*, where qualitative research uses a probabilistic sampling frame that survey researchers can employ to test ethnographic concepts (Sánchez-Jankowski, 1992; Small et al., 2008). For example, in their study of childcare centers in New York City, Small and colleagues (2008) begin by interviewing personnel in 23 childcare centers within four types of neighborhood, with each neighborhood spanning three to five census tracts. Their quantitative analysis then tested the narratives and themes that emerged from qualitative fieldwork on a random sample of 293 (out of 1683) childcare centers in 243 census tracts. This design follows *prospective sampling* because the qualitative interviews from 23 childcare centers were selected randomly across multiple census tracts and then tested quantitatively on a random sample of childcare centers that spanned several hundred census tracts. Thus, the probability of selection, which can be derived mathematically, was either wholly or partially aligned in both the qualitative and quantitative units of their study.

Second, using *retrospective sampling*, qualitative researchers can draw on probabilistic sampling frames from existing survey research or a population list, if

Table 5. Odds ratios from a logistic regression predicting the zero[th] (or negative) case in narrative saturation.

	All Observations		Non-Working Poor		Working Poor		Total Changes	
	Unaligned	*Aligned*	*Unaligned*	*Aligned*	*Unaligned*	*Aligned*	*Direction*	*Significance*
NH-Black	0.99	1.29	0.96	1.36	1.17	0.96	3/3	0/3
	(0.09)	(0.25)	(0.09)	(0.30)	(0.27)	(0.38)		
Hispanic	0.99	1.06	0.99	1.14	0.98	0.79	2/3	0/3
	(0.07)	(0.17)	(0.08)	(0.21)	(0.21)	(0.29)		
NH-Other	1.11	1.17	1.12	1.09	1.10	1.75	0/3	0/3
	(0.09)	(0.18)	(0.09)	(0.16)	(0.29)	(1.05)		
Parental Incarceration	1.01	1.00	1.04	1.17	**0.79**	**0.30***	0/3	1/3
	(0.14)	(0.23)	(0.15)	(0.28)	(0.32)	(0.16)		
NH-Black x Parental Incarceration	0.72	0.85	0.79	0.83	0.55	1.90	1/3	0/3
	(0.24)	(0.42)	(0.28)	(0.45)	(0.46)	(1.99)		
Hispanic x Parental Incarceration	0.77	0.96	0.85	0.95	0.39	0.48	0/3	0/3
	(0.24)	(0.55)	(0.28)	(0.56)	(0.43)	(0.57)		
NH-Other x Parental Incarceration	0.79	1.01	0.73	1.16	1.14	0.47	3/3	0/3
	(0.24)	(0.59)	(0.25)	(0.71)	(0.85)	(0.49)		
Number of Directionality Changes (underlined)	3/7		3/7		3/7		9/21	—
Number of Significance Changes (bolded)	0/7		0/7		1/7		—	1/21

^p < .10 *p < . 05 **p < . 01 ***p < .001

Source: Authors' calculation from National Survey of Children's Health (NSCH) and Monte Carlo simulated data.

Note: Non-Hispanic whites are the baseline group.

available, in order to locate their field sites and experiences within a broader distribution of sampled units or the target population. In turn, this allows quantitative researchers who may derive and test hypotheses from ethnographic findings to calculate weights either from a population stratified sample or from weights already established in national or local surveys. Pearce's (2002) study of religion's influence on childbearing preferences is an example of retrospective sampling aligned across sequential phases of survey and ethnographic research in the same study. Similarly, in his study of the lindy hop, Hancock (2005, 2013) traveled to Stockholm, Herräng, Copenhagen, Toronto, and Montreal to examine if the narratives and themes generated from his observations in Chicago aligned with international lindy hop dance camps. In this study, the full population list of lindy hop dance camps (15 annually) and the number of instructors and dancers could be estimated. By venturing to a third of those camps, one of which (Herräng) included approximately 4000 instructors and students from at least 50 countries, he *retrospectively sampled* from a population list of sites. Thus, even if Hancock's retrospective sampling was inadvertent, this process allows any future quantitative researcher to derive sampling weights (including stratified sampling weights) associated with his field sites. Importantly, a retrospective sampling technique can allow for separate studies to be conducted by different researchers, which has the potential to inform and improve theory construction and model fit for a particular study. However, qualitative researchers revisiting respondents included in national or local surveys would need to obtain IRB approval, respondent consent for future contact, and support from survey administrators (see Leahey, 2007).

Finally, in *aligned concurrent sampling*, qualitative and quantitative researchers sample the same units from the underlying population distribution. Whereas *prospective* and *retrospective sampling* may require calculating statistical adjustments to sampling weights in order to align qualitative and quantitative samples, no adjustments are necessary under *aligned concurrent sampling* because the weights are implicitly the same given the way the qualitative and quantitative units are sampled. Qualitative researchers should simultaneously embed themselves in the quantitative research process by gaining access to field sites and respondents as survey data are being collected. One drawback to this solution is timing; surveys take a long time to develop, pretest, and administer, and qualitative researchers will need to obtain IRB approval and respondent consent to conduct supplemental field observations and interviews on respondents and neighborhoods.

Conclusion and implications

Ongoing debates within the field of ethnography have flared against the backdrop of mass incarceration, as social inequality research takes a mixed-methods turn. Sampling and case selection procedures, while largely overlooked, are at the root of these controversies. The blending of qualitative and quantitative data in dual research designs offers great promise in expanding and clarifying our understanding of the processes and mechanisms of racial inequality and mass incarceration,

but it also raises critical questions about scientific validity and the limits of generalizability in such research. Any empirical assessment of these questions must contend with the distinct epistemological and ontological assumptions that underlie ethnographic and quantitative methods of inquiry, as well as the fact that even the most rigorous social scientific methods proceed on premises that can often only approximate unknown social distributions in the real world. For this reason, we have used a combination of simulated and real world data to examine these questions in a quasi-fictional social reality where the number of observations necessary for reaching a saturated knowledge, as well as the distributions of narratives derived from field observations, are known to researchers. Even in this scientific fantasyland, we demonstrate how the introduction of bias at the sampling stage of qualitative research – when cases are non-randomly selected through convenience, snowball or purposive sampling – leads to erroneous conclusions about the relationships between race and parental incarceration. One can only imagine the scale of bias, then, that exists in field research where assumptions are not explicated about the distributional properties of cases and phenomena under study.

Moreover, our results show that the exclusion of negative and disconfirming cases alters both the direction and level of significance associated with demographic groups that routinely experience social inequality. Failure to address non-response, zeroth cases, and the non-random sampling of observations and field sites exacerbates bias in studies, particularly if this work contains low internal validity or cannot be replicated because of spatial and temporal changes. Conclusions drawn and then generalized to out-of-sample populations under the guise of 'pragmatically' or 'synergistically' borrowing the independent strengths from qualitative and quantitative methods may contain misspecified causal relationships because the weaknesses of each method are also embedded in the inner workings of mixed-methods research designs.

Aligning the qualitative and quantitative population distribution is vital to ensuring internally and externally valid findings in mixed-methods research. While our work has showcased the limits of generalizability in mixed-methods studies on parental incarceration (as a heuristic) when the sampling procedures in dual research designs are not aligned, a parallel (albeit distinct) concern exists in quantitative studies on social inequality. Recent scholarship highlights the consequences of drawing conclusions about social inequality using household-based surveys. Because inmates are institutionalized and excluded from household-based sample surveys, classic measures of social inequality are biased (Pettit, 2012), necessitating the integration of multiple data sources to re-estimate indicators of stratification by race. Estimates of racial inequality in educational completion, employment, wages, voter turnout, wealth, and health metrics have been shown to be biased because inmates are not included in conventional surveys (Sykes and Maroto, 2016; Pettit and Sykes, 2015; Ewert et al., 2014; Sykes and Pettit, 2014; Pettit, 2012; Western and Pettit, 2005). The elision of inmates from national data sources routinely used to measure racial progress in America means that 'we

continue to live in an age of de facto discrimination' despite civil rights legislation and social policies aimed at reducing and redressing social inequality (Pettit and Sykes, 2015: 608). Mixed-methods studies, then, must account for existing sampling limitations in both qualitative and quantitative studies if inferences are to be valid for the assessment of legal reforms and the construction of social policy.

Our argument for the alignment of sampling procedures and statistical weights across methodologies in certain mixed-methods research designs does not imply the *assimilation* of qualitative inquiry into positivistic quantitative paradigms. Rather, the implication is that there are limits to the generalizability of findings and the value added by MMR when sampling procedures are not consistent across quantitative and qualitative research strands. If the research objective of a MMR study is to achieve the greatest breadth-depth optimization possible from blending methods to discover findings that are generalizable to a population or site beyond the cases examined, then the research has already begun to proceed from an ontological and epistemological position premised on some form of probability-based sampling. Failure to recognize probabilistic research questions as such, in addition to the sampling procedures such questions necessitate, mask a form of 'assimilation' that has, in fact, already taken place and has grave consequences for the findings of a mixed-methods study.

Far from being insularly academic, these debates have a bearing on how social problems and pathologies at the intersection of race, law, and inequality are conceived by the public-at-large and the policymakers who, regardless of the researcher's intention, may be moved to intervene in the lives and places under scientific gaze. A thorny implication for race, law, and social inequality research is, thus, that even when a study is *not* designed to achieve ideal breadth-depth optimization or generalizability, it may be interpreted as such when translated into policy arenas.

The burgeoning use of MMR in social science research has the capacity to draw policymakers and the general public into debates about social policy (see e.g. Leahey, 2007; Pearce, 2007, 2012). However, mixed-methods findings are not in the public's best interest if qualitative research is not internally valid (or consistent) through reproducibility, and quantitative research drawing on qualitative insights is not externally valid (or generalizable to the population) due to the use of non-probability based samples to test theoretical processes and mechanisms included in statistical models. Policymakers may 'like causal models that appear to support their position on important questions,' but social scientists must acknowledge and attend to the limits of research, whether mixed-methods or not, for 'it is the continued misuse of the models by scholars that makes the process respectable' (Zuberi, 2000: 182).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. For more detail on the data see Sykes and Pettit (2014, 2015).
2. Because the same themes may emerge across repeated interviews or field observations for a case, assumptions of statistical independence may be violated if all themes for a respondent are included in regression analyses without clustering on individual or case IDs. However, we test the most extreme version, which is that themes are completely independent of each other. For this reason, we code the most prevalent theme for all possible observations within a particular case. Theoretically, the statistical dependence between themes within a case is not the same as statistical dependence for the most prevalent theme between cases. Given that we coded the most prevalent theme under independence within a case, and then conduct our analysis on the most prevalent theme between cases, we do not violate assumptions of independence in our hypothesis tests because the most prevalent theme for a case is independent from one case to the other. Furthermore, if these themes were correlated within and between cases, the overall distribution of themes would appear more uniform (see Table 1).
3. These five thematic narratives can be extrapolated to represent different ideal types such as neighborhoods, institutions, or particular social classes. For our purposes, however, we concentrate solely on themes that may represent mechanisms of racial inequality in parental incarceration (e.g. intergenerational transmission of incarceration; unobserved labor market factors; educational levels; mental health characteristics; and differential surveillance).
4. The simulation was conducted in R, and the seed was set at 1001. Although Monte Carlo simulations are normally repeated 1000, 10,000, or 100,000 times, we conducted one simulation for two reasons. First, realistically, ethnographic studies are largely conducted once and are not repeated. Second, repeating the simulation upwards of 100,000 times may produce bias for all quantitative measures in our study, which would prevent us from establishing a minimum existence proof of the problem in mixed-methods research where an investigator only conducts his or her study once. However, under repeated sampling and estimation, one could learn how often a particular mixed-methods study is biased for the measures contained within the analysis – which is important in its own right – but this line of inquiry is not the immediate aim of our article.
5. From our perspective, a non-response observation and an unobserved case both lack the capacity to reach saturation, but for different reasons, resulting in zeros from the Poisson model. Practically, the non-response and unobserved case both suffer from a complementary but similar problem: neither has a thematic narrative associated with them, preventing both forms of 'missingness' from ever reaching saturation. The different reasons do not affect our substantive findings, with respect to the effect of bias in observed thematic narratives within a mixed method study.
6. Solon, Haider and Wooldridge (2015) show the conditions wherein weights should be used to achieve more precise estimates including correcting for heteroscedasticity, obtaining consistent estimates by correcting for endogenous sampling, and identifying average partial effects in the presence of unmodeled heterogeneity of effects. Further, they argue that 'a practical question is not whether a chosen [model] specification is

exactly the true datagenerating process, but rather whether it is a good enough approximation to enable nearly unbiased and consistent estimation of the causal effect of interest' (p. 304). Solon, Haider, and Wooldridge recommend reporting both weighted and unweighted estimates (see also Dickens (1990) on this point). We contend that comparing weighted (aligned) and unweighted (unaligned) estimates is one method for assessing bias in coefficients due to differences in sampling approaches within a mixed-method study.

7. We also investigated whether findings in Table 2 hold using a Poisson regression model. Estimates from incident rate ratios display an identical pattern to statistics reported from the OLS model. Given our audience, we have chosen to present the OLS model for a more accessible interpretation of explained variation (R^2) and model selection over reporting AIC or BIC estimates from the Poisson model; however, both sets of findings are available upon request.
8. We use multinomial logit regression because our outcome consists of five possible themes. For a fuller description of this method, see Long (1997).

References

Abbott A (2004) *Methods of Discovery: Heuristics for the Social Sciences*. New York: W.W. Norton & Company.

Blalock HM Jr (1968) The measurement problem: A gap between languages of theory and research. In: Blalock HM Jr and Blalock AB (eds) *Methodology in Social Research*. New York: McGraw-Hill, pp. 5–27.

Burawoy M (2005) 2004 Presidential address: For public sociology. *American Sociological Review* 70: 4–28.

Cohen PN (2015) Survey and ethnography: Comment on Goffman's 'On the Run', 28 June. Available at: https://familyinequality.wordpress.com/2015/05/28/on-the-ropes-goffman-review/ (accessed July 2017).

Creswell JW, Plano Clark VL, Gutmann M, et al. (2003) Advanced mixed methods research designs. In: Tashakkori A and Teddlie C (eds) *Handbook of Mixed Methods in Social and Behavioral Research*. Thousand Oaks, CA: SAGE, pp. 209–240.

Desmond M (2014) Relational ethnography. *Theory and Society* 43: 547–579.

Diaconis P (2009) The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society* 46(2): 179–205.

Dickens WT (1990) Error components in grouped data: Is it ever worth weighting? *Review of Economics and Statistics* 72(2): 328–33.

Duneier M (1999) *Sidewalk*. New York: Farrar, Strauss and Giroux.

Duneier M (2004) Scrutinizing the heat. *Contemporary Sociology* 33: 139–50.

Duneier M (2006) Research note: Ethnography, the ecological fallacy, and the 1995 Chicago heat wave. *American Sociological Review* 71: 679–688.

Duneier M (2011) How not to lie with ethnography. *Sociological Methodology* 41: 1–11.

Evans M, Hastings N and Peacock B (2000) *Statistical Distributions*, 3rd ed. New York: Wiley-Interscience.

Ewert S, Sykes BL and Pettit B (2014) The degree of disadvantage: Incarceration and inequality in education. *The Annals of the American Academy of Political and Social Science* 651(1): 24–43.

Gill J (2002) *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: Chapman & Hall/CRC.

Goffman A (2014) *On the Run: Fugitive Life in an American City*. Chicago: University of Chicago Press.
Greene JC (2007) *Mixing Methods in Social Inquiry*. San Francisco: Jossey-Bass.
Guba EG and Lincoln YS (2005) Paradigmatic controversies, contradictions, and emerging confluences. In: Denzin NK and Lincoln YS (eds) *The SAGE Handbook of Qualitative Research*. Thousand Oaks, CA: SAGE, pp. 191–215.
Hall B and Howard K (2008) A synergistic approach: Conducting mixed methods research with typological and design considerations. *Journal of Mixed Methods Research* 2(3): 248–269.
Hancock BH (2013) *American Allegory: Lindy Hop and the Racial Imagination*. Chicago: University of Chicago Press.
Hancock BH (2005) Steppin' out of whiteness. *Ethnography* 6(4): 427–461.
Haraway D (1988) Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies* 14(3): 575–599.
Hesse-Biber S (2015) Mixed methods research: The 'thing-ness' problem. *Qualitative Health Research* 25(6): 775–788.
Katz J (1997) Ethnography's warrants. *Sociological Methods and Research* 25: 391–423.
Kemper E, Stringfield S and Teddlie C (2003) Mixed methods sampling strategies in social science research. In: Tashakkori A and Teddlie C (eds) *Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks, CA: SAGE, pp. 273–296.
Klinenberg E (2002) *Heat Wave: A Social Autopsy of Disaster in Chicago*. Chicago: University of Chicago Press.
Klinenberg E (2004) Overheated. *Contemporary Sociology* 33(5): 521–528.
Klinenberg E (2006) Blaming the victims: Hearsay, labeling, and the hazards of quick-hit disaster ethnography. *American Sociological Review* 71(4): 689–698.
Johnson J and DiNardo J (1997) *Econometric Methods*, 4th ed. New York: McGraw Hill.
Lamont M and Swidler A (2014) Methodological pluralism and the possibilities and limits of interviewing. *Qualitative Sociology* 37(2): 153–171.
Lamont M and White P (2009) Workshop on Interdisciplinary Standards for Systematic Qualitative Research. Washington, DC: National Science Foundation. Available at: http://www.nsf.gov/sbe/ses/soc/ISSQR_workshop_rpt.pdf (accessed July 2017).
Leahey E (2007) Convergence *and* confidentiality? Limits to the implementation of mixed methodology. *Social Science Research* 36: 149–158.
Lewis-Kraus G (2016) The trials of Alice Goffman. *New York Times Magazine*, 12 January. Available at: http://www.nytimes.com/2016/01/17/magazine/the-trials-of-alice-goffman.html?_r=0 (accessed July 2017).
Long JS (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: SAGE.
Lubet S (2015) Ethics on the run. *The New Rambler*. Available at: http://newramblerreview.com/book-reviews/law/ethics-on-the-run (accessed July 2017).
Lucas SR (2014) Beyond the existence proof: Ontological conditions, epistemological implications, and in-depth interview research. *Quality & Quantity* 48(1): 387–408.
Lucas SR and Szatrowski A (2014) Qualitative comparative analysis in critical perspective. *Sociological Methodology* 44: 1–79.
O'Cathain A and Thomas K (2006) Combining qualitative and quantitative methods. In: Pope C and Mays N (eds) *Qualitative Research in Health Care*, 3rd ed. Oxford: Blackwell, pp. 102–111.

Pearce LD (2002) Integrating survey and ethnographic methods for systematic anomalous case analysis. *Sociological Methodology* 32: 103–132.

Pettit B (2012) *Invisible Men: Mass Incarceration and the Myth of Black Progress*. New York: Russell Sage Foundation.

Pettit B and Sykes BL (2015) Civil rights legislation and legalized exclusion: Mass incarceration and the masking of inequality. *Sociological Forum* 30(S1): 589–611.

Platt T (2016) *On the run from her critics: Alice Goffman's ethnography*. Available at: http://goodtogo.typepad.com/tony_platt_goodtogo/2016/02/index.html (accessed 19 May 2017).

Ralph L (2015) The limitations of a 'dirty' world: A review of Alice Goffman's *On the Run* and Victor M. Rios' *Punished*. *DuBois Review* 12(2): 441–451.

Rios V (2015) Review: *On the Run: Fugitive Life in an American City*, by Alice Goffman. *American Journal of Sociology* 121(1): 306–308.

Sánchez-Jankowski M (1992) *Islands in the Street: Gangs and American Urban Society*. Berkeley, CA: University of California Press.

Sánchez-Jankowski M (2002) Representation, responsibility and reliability in participant-observation. In: May T (ed.) *Qualitative Research in Action*. London: SAGE, pp. 144–160.

Sharkey P (2015) *On the Run: Fugitive Life in an American City*, by Alice Goffman. *Social Service Review* 89(2): 407–412.

Small ML (2009) How many cases do I need?: On science and the logic of case selection in field-based research. *Ethnography* 10(1): 5–38.

Small ML (2011) How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology* 37: 57–86.

Small ML (2013) Causal thinking and ethnographic research. *American Journal of Sociology* 119(3): 597–601.

Small ML, Jacobs E and Massengill R (2008) Why organizational ties matter for neighborhood effects: A study of resource access through childcare centers. *Social Forces* 87: 387–414.

Solon G, Haider SJ and Wooldridge J (2015) What are we weighting for? *The Journal of Human Resources* 50(2): 301–316.

Sykes BL and Pettit B (2015) Severe deprivation and system inclusion among children of incarcerated parents in the United States after the Great Recession. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 1(2): 108–132.

Sykes BL and Maroto M (2016) A wealth of inequalities: Mass incarceration, employment, and racial disparities in household wealth, U.S. 1996–2011. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 2(6): 129–152.

Sykes BL and Pettit B (2014) Mass incarceration, family complexity, and the reproduction of childhood disadvantage. *The Annals of the American Academy of Political and Social Science* 654: 127–149.

Tashakkori A and Teddlie C (eds) (2003) *Handbook of Mixed Methods in Social and Behavioral Research*. Thousand Oaks, CA: SAGE.

Tavory I and Timmermans S (2013) A pragmatist approach to causality in ethnography. *American Journal of Sociology* 119(3): 682–714.

Teddlie C and Yu F (2007) Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research* 1(1): 77–100.

Turney K (2014) Stress proliferation across generations? Examining the relationship between parental incarceration and childhood health. *Journal of Health and Social Behavior* 55: 302–319.
Venkatesh S (2013) The reflexive turn: The rise of first-person ethnography. *Sociological Quarterly* 54(1): 2–8.
Wacquant L (2002) Scrutinizing the street: Poverty, morality, and the pitfalls of urban ethnography. *American Journal of Sociology* 107(6): 1468–1532.
Western B and Pettit B (2005) Black-white wage inequality, employment rates, and incarceration. *American Journal of Sociology* 111: 553–578.
Western B, Braga A and Kohl (2014) A longitudinal survey of newly-released prisoners: Methods and design of the Boston Reentry Study. Unpublished manuscript. Available at http://scholar.harvard.edu/files/brucewestern/files/brs_research_design.pdf?m=1418851307 (accessed July 2017).
Wilson WJ (2014) The travails of urban research. *Contemporary Sociology: A Journal of Reviews* 43: 824–28.
Zuberi T (2000) Deracializing social statistics: Problems in the quantification of race. *The Annals of the American Academy of Political and Social Science* 568: 172–185.

Author Biographies

Bryan L Sykes is an assistant professor of Criminology, Law and Society (and Sociology and Public Health, by courtesy) at the University of California-Irvine. He holds a Joint PhD in Demography and Sociology from the University of California-Berkeley. His research focuses on demography, mass incarceration, criminology/deviance, population health, and research methodology. Professor Sykes' research has appeared in *The Lancet*, *JAMA*, *Medicine*, *The Russell Sage Foundation Journal of the Social Sciences*, *The Annals of the American Academy of Political and Social Science, Sociological Forum, and Crime & Delinquency*, among others, and is forthcoming in *Sociological Perspectives* and *the Annual Review of Criminology*. He is currently collaborating on a multi-state mixed-method data collection effort to assess the legal history and social consequences of monetary sanctions across different jurisdictions within the United States.

Anjuli C Verma is a Chancellor's postdoctoral fellow in Jurisprudence and Social Policy at the University of California, Berkeley. She holds a PhD in Criminology, Law and Society from the University of California, Irvine. Her research examines punishment, law, and inequality from an interdisciplinary perspective using multiple methods. Anjuli's work appears in *Law & Society Review*, *The Annals of the American Academy of Political and Social Science*, *The Oxford Handbook on Prisons and Imprisonment*, *The British Journal of Criminology*, *The American Journal of Bioethics* and is forthcoming in *Sociological Perspectives* and *The Oxford Research Encyclopedia of Criminology and Criminal Justice*. She is currently collaborating on a multi-state mixed-methods study of monetary sanctions in the US criminal justice system.

Black Hawk Hancock is an associate professor of Sociology at DePaul University. He received his PhD in Sociology from the University of Wisconsin at Madison. He is an ethnographer whose work focuses on issues of race and culture. His first ethnographic monograph, *American Allegory: Lindy Hop and the Racial Imagination*, was published by The University of Chicago Press in 2013. His second book, *In-Between Worlds: Mexican Kitchen Workers in Chicago's Restaurant Industry*, is currently under contract with The University of Chicago Press.