Published in final edited form as:

Sociol Methods Res. 2014 February ; 43(1): 137–170. doi:10.1177/0049124113506407.

Local Dependence in Latent Class Analysis of Rare and Sensitive Events

Marcus E. Berzofsky¹, Paul P. Biemer^{1,2}, William D. Kalsbeek²

¹RTI International, Research Triangle Park, NC, USA

²The University of North Carolina, Chapel Hill, NC, USA

Abstract

For survey methodologists, latent class analysis (LCA) is a powerful tool for assessing the measurement error in survey questions, evaluating survey methods, and estimating the bias in estimates of population prevalence. LCA can be used when gold standard measurements are not available and applied to essentially any set of indicators that meet certain criteria for identifiability. LCA offers quality inference, provided the key threat to model validity—namely, local dependence—can be appropriately addressed either in the study design or in the model-building process. Three potential causes threaten local independence: bivocality, behaviorally correlated error, and latent heterogeneity. In this article, these threats are examined separately to obtain insights regarding (a) questionnaire designs that reduce local dependence, (b) the effects of local dependence on parameter estimation, and (c) modeling strategies to mitigate these effects in statistical inference. The article focuses primarily on the analysis of rare and sensitivity outcomes and proposes a practical approach for diagnosing and mitigating model failures. The proposed approach is empirically tested using real data from a national survey of inmate sexual abuse where measurement errors are a serious concern. Our findings suggest that the proposed modeling strategy was successful in reducing local dependence bias in the estimates, but its success varied by the quality of the indicators available for analysis. With only three indicators, the biasing effects of local dependence can usually be reduced but not always to acceptable levels.

Keywords

National Inmate Survey (NIS); local dependence; correlated error; bivocality; latent heterogeneity; model misspecification; expeculation

Introduction

The survey methods research literature defines an optimal survey design as one that strikes an optimal balance between survey costs and error (Groves 1989; Biemer and Lyberg 2003). The primary sources of survey error are the survey frame (primarily coverage errors), the

Corresponding Author: Marcus E. Berzofsky, RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709, USA., berzofsky@rti.org.

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.

sampling process, nonresponse, and measurement error. The focus of this article is on the latter error source and considers important issues that arise in the evaluation of measurement errors using latent class analysis (LCA).

Measurement errors arise as a result of poorly designed survey questions, interviewers, respondents, data collection modes (e.g., in person, telephone, web, etc.), the survey setting (e.g., private or nonprivate), and many other factors (Biemer and Lyberg 2003). Quantifying the effects of measurement errors in the estimation process is the first step toward understanding the root causes of the errors and ultimately reducing the errors in future implementations of a survey. Additionally, knowledge of the distributional properties of measurement errors is useful for compensating for their effects in data analyses and for interpreting analytical results.

In social surveys, many questions are close-ended giving rise to categorical responses. Continuous measures such as age and income are often discretized and treated as categorical in an analysis. For categorical data, the main consequence of measurement error at the item level is misclassification which leads to measurement bias and variance in survey estimation (Biemer 2011). An important modeling approach for studying classification error is LCA (Lazarsfeld and Henry 1968). Although LCA is more than 50 years old, it has only recently been used to study the sources of measurement error in surveys (Biemer 2004, 2011). LCA can be very effective for quantifying classification error in survey data as long as the model assumptions are reasonably satisfied. But like any modeling approach, using it improperly can lead to invalid results. Unfortunately, the proper use of latent class models is challenging for many data analysts because, in survey situations, the modeling process can be quite complex. Moreover, practical guidance on key issues is either unavailable or is too general to be useful in specific situations. Modeling strategies often differ depending upon the purpose of the analysis, the type of data being analyzed, and the types of models being fit.

This article focuses specifically on the analysis of rare and sensitive events such as illicit drug use, deviant sexual behavior, and some types of crime victimizations where the goal of the analysis is to evaluate the classification error in the survey responses. For example, in the National Crime Victimization Survey, there is much concern that rape, sexual assault, and other major personal crimes go unreported (see, e.g., Groves and Cork 2008). LCA is a tool that can help develop (a) clear and concise definitions of sensitive events, (b) an appropriate contextual framework for asking about them in an interview, and (c) questioning methods that appropriately convey the proper meanings so that both underreporting and overreporting are minimized. LCA can also aid in identifying population and geographic domains that are most prone to misclassification. Such results provide insights about the causes of the error so that interview procedures that elicit more accurate responses can be developed for these subpopulations. Of course, producing these benefits requires valid results from the LCA. Thus, the goal of this article is to provide guidance to LCA analysts, particularly those focusing on rare and sensitive events, on how to address the central assumption of any LCA—namely, local independence.

The context for the article is studies where only a few indicators of the latent variable are available and interest is on studying the error properties of at least one of the indicator.

Designs having only three indicators of the latent variable, the minimum required for identifiability, are considered in some detail. For example, the researcher may be interested in conducting an LCA to evaluate the error in a survey item that asks about abortions. Two additional indicators of abortions are created and embedded in the survey questionnaire. Such evaluation designs are quite common in survey methods applications for reasons described in the next section.

The next section also provides a review of the problem and the general approach adopted in the article. The third section provides some simulation results demonstrating the nature of each cause of local dependence bias, particularly for rare and sensitive events, how each cause of local dependence manifests itself as estimation bias, and factors that determine the severity of the bias. Various approaches for modeling local dependence and their effectiveness for each source of local dependence are also reviewed. In the fourth section, the tactics described in the third section are combined in a strategy for modeling local dependence. This strategy is then applied to data on rape and sexual assault from a national survey of prison inmates. Finally, the fifth section summarizes the findings and forms conclusions regarding latent class modeling for rare and sensitive events.

Background

LCA and Survey Error

Survey methodologists are often faced with a number of survey design alternatives—that is, question methods, modes of interview, reference periods, and so on—where the comparative costs and error implications of each alternative determine the final choice. Such situations are characterized by the absence of a valid gold standard, a very limited set of evaluative data, and research objectives having considerable scope (e.g., many disparate questionnaire items are to be evaluated). Often, embedding three indicators of the same construct in the questionnaire is practically infeasible.

For example, when redesigning a national survey of drug abuse, methodologists may be interested in numerous questions that deal with many different types of drug use. The usual approach for evaluating the error for so many items is to conduct a reinterview survey—that is, a survey of original survey respondents that seeks to obtain replicate measurements of the same survey items. This approach may be combined with embedding replicate measurements of some survey items in one of the interviews to increase the number of indicators for analysis. However, both approaches are problematic. Reinterview surveys are an expensive option for collecting repeated measurements since they essentially double the interviewing costs. Further, in order to avoid temporal confounding (which is a violation of local independence), the reinterview survey should be conducted within just a few weeks of the original interview (Bailar 1968; Biemer and Lyberg 2003). Further, survey respondents are unlikely to cooperate beyond one repetition of the same interview, especially within this short time period. In addition, reinterviewing too soon after the original interview can lead to memory effects leading to correlated errors (Bailar 1968)—another cause of local dependence.

Embedded replication is also problematic because lengthening the interview in this way creates respondent burden and adds costs that are difficult to justify for production surveys. In addition, alternative question wordings about the same event must be constructed to conceal the redundancy; otherwise, respondents may either decline to respond or force responses to agree across the repeated questions, a form of satisficing (Krosnick and Alwin 1987; Krosnick 1991). However, rewording the questions for this purpose can lead to variations in the constructs being measured which can also violate local independence.

Rather than embedding replicate measurements, it is also possible to obtain one or more repeated measurements from administrative data sources; for example, data on crime victimizations could be obtained from official crime reports. Administrative records are available for only a very limited set of characteristics and are usually not a viable option, especially for rare and sensitive event. In addition, these data are usually plagued by definitional and coverage issues. As an example, official crime data only cover reported victimizations and definitions often vary by region. Definitional differences between the record and the survey pose additional concerns for local independence.

Thus, it is often quite difficult to collect more than three indicators of a construct in survey error evaluations which is the minimum number of indicators required for identifiability of a latent class model. This article provides specific guidance for this situation although the lessons learned from our consideration of three indicator models can be easily extended to models with four or more indicators. The three measurements may be obtained in any manner: for example, from the original survey and two reinterview surveys, three replicate measurements embedded in the original survey, two replicate measurements and either a reinterview or an administrative record data value, or any other configuration.

Local Independence

Let A , B , and C denote the three dichotomous indicators of a well-defined dichotomous, latent construct denoted by X . Polytomous indicators can be handled by straightforward extensions of the dichotomous indicator methods. For evaluating survey error, X represents the true value of the outcome variable and A , B , and C represent the survey realizations of X , that is, the indicators of X . For example, for evaluating questions regarding past year cocaine use, X is the true status of an individual's past-year use, A is a determination of X based upon the survey, and B and C may be replicate measurements of X that were embedded in the questionnaire for the sole purpose of evaluating A using LCA. Using conventional LCA notation, let π_a^A , π_{ab}^{AB} , and π_{abc}^{ABC} denote marginal probabilities $\Pr(A = a)$, $\Pr(A = a, B = b)$, $\Pr(A = a, B = b, C = c)$, respectively, $\pi_{a|x}^{A|X}$, $\pi_{b|xa}^{B|XA}$, and $\pi_{c|xab}^{C|XAB}$ denote the conditional probabilities $\Pr(A = a|X = x)$, $\Pr(B = b|X = x, A = a)$, $\Pr(C = c|X = x, A = a, B = b)$, respectively, with other marginal and conditional probabilities defined analogously.

The indicators A , B , and C are *locally independent indicators of X* if and only if

$$\pi_{abc|x}^{ABC|X} = \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}. \quad (1)$$

If equation (1) does not hold, then A , B , and C are called *locally dependent*. Essentially, local independence means that the X fully explains the relationship among indicators in that, after accounting for X , A , B , and C are mutually independent. Local independence is the key assumption underlying the standard latent class model and, as we show in the next section, its failure can severely bias the standard model parameters. Unfortunately, the assumption seldom holds in practice (Pepe and Janes 2007; Sepulveda, Vicente-Villardón, and Galindo 2008) and the standard model must be extended to obtain valid inference. These so-called *extended latent class models*, first proposed by (Hagenaars 1988), will be discussed in the third section. This article will focus on the root causes of local dependence, how they manifest themselves in an LCA, approaches for diagnosing local dependence, and a systematic process for appropriately and parsimoniously extending the standard model to properly account for each cause separately or in combination.

As shown in Biemer (2011), the violations of local independence stem from three essential causes: *bivocality*, *behaviorally correlated error*, and *latent heterogeneity*. *Bivocality* arises when the indicators do not measure a common latent variable, X . So, for example, A is an indicator for X , while B is an indicator for Y and $\Pr(X = Y)$ is not unity (see, e.g., Wiggins 1973; Alwin 2007). To illustrate, suppose A indicates “past-year cocaine use” while B indicates “past-year illegal drug use.” A and B are bivocal indicators because an individual may truthfully answer *yes* to B (i.e., past-year illegal drug use) and *no* to A (i.e., no past-year cocaine use). Hence, $\Pr(X = Y)$ is not unity. Further examples of bivocality can be found in Biemer (2011).

Behaviorally correlated error arises when a respondent’s response to B is influenced by having been asked about A , or vice versa. Suppose A is asked before B in an interview and the respondent intentionally falsifies A . A behaviorally correlated error results if the respondent merely repeats error (which may have been a guess) in response to B . The risk of this occurring is particularly high in embedded replication designs where A and B are observed during the same interview. Similarly, respondents may misunderstand some common concept underlying both A and B and, consequently, may respond with the same erroneously response to both whether in the same interview or two separate interviews.

Finally, *latent heterogeneity* arises when error probabilities vary in the population by an unknown grouping variable (or, domain indicator variable). For example, reports of excess alcohol consumption might be falsified at higher rates for persons of greater socioeconomic status due to increased sensitivity to social undesirable questions. It is also known that poorly educated respondents have higher risks of misclassification because they misinterpret the questions more often (see, e.g., Tourangeau, Rips, and Rasinski 2000). Such variations in the response probabilities across population domains will also induce dependencies among the indicators that are not explained by X , thus violating equation (1). If the domains creating these variations are known, grouping variables for the domains can be added to the model to explain the variation. However, there will always be unknown causes of heterogeneity that can be explained only approximately by the available data.

We will show that each of these three conditions will lead to a violation of local independence and, consequently, biased inference. The magnitudes of the biases may vary

across the parameters with some parameters being considerably biased while other parameters exhibiting very little bias (see also Sepulveda et al. 2008). Likewise, some indicators may be more affected by local dependence than others and the choice of indicators is a key consideration for bias mitigation.

With only three indicators, it is impossible in practical situations to distinguish among the three causes of local independence in an analysis. Nevertheless, a fundamental finding of our research, and the foundation of the proposed modeling approach, is that there are distinct advantages in considering how the three underlying causes of local dependence can be mitigated in the design and analysis stages.

At the design stage, the risks of false inference in an LCA can be greatly reduced by developing research designs that appropriately consider each cause of local dependence. Bivocality can be prevented by developing questions (perhaps through cognitive testing) that are ostensibly univocal. This may be quite challenging, particularly as the number of indicators increases, due to the risk of greater behavioral correlated error. For example, there are only a few ways to ask about past-year cocaine use or rape victimizations. Embedding multiple indicators of these outcomes in the same questionnaire may require deliberately creating bivocal measurements to obscure the redundancy. Otherwise, if respondents become aware of the redundancy, they may repeat early responses from memory (a form of satisficing) inducing greater behavioral correlation. Additional guidance for dealing with bivocality in designing questionnaires will be provided subsequently.

Reducing the effects of behavioral correlated error is also quite challenging to research designers. For example, the timing of reinterview surveys should be such that respondents are unable to merely recall their prior responses, yet they are still able to access the same reference period as they were in the original survey. Time-reference recall errors may result if the reference period is held fixed; however, bivocality may be induced if the reference period is allowed to adjust (or float) according to the timing of the reinterview.

Finally, as we will show, latent heterogeneity can more effectively be mitigated during data analysis if appropriate, theoretically vetted grouping variables have been collected in the survey. This may require prior research on the causes of misclassification so that the most effective covariates mitigating latent heterogeneity can be collected in the survey.

However, an ideal design is both illusive and often impractical. In routine survey work, measurement error evaluation is often of secondary importance to the production of population statistics and the later objective usually dominates the survey design. This reality highlights the importance of using analytical methods that effectively address each potential cause of local dependence, particularly for situations where the number of indicators is minimal. Thus, our article demonstrates how each of the three causes of local dependence affects estimation, particularly for three-indicator models, and proposes a modeling strategy for reducing their effects that systematically addresses each cause.

As previously noted, our focus is on *rare* events—that is, events with low prevalence (say, 10 percent or less)—and *sensitive* questions—that is, questions about behaviors that are often underreported (high false-negative probabilities) and seldom overreported (very small false-

positive probabilities). In the crime victimization literature, false-negative probabilities can attain fairly high levels—say, 30 to 40 percent or more, while false positives may only be 1 percent or less (see, e.g., Groves 1989). Such events are central to the study of socially undesirable phenomenon, yet there has been little research on methods for evaluating errors in the measurement of these phenomena. Examples including measuring and reducing measurement error in surveys of illegal activities, socially undesirable behaviors, rare diseases, and so on are critical for establishing social policies in these areas.

Within this framework, the remainder of the article will explore three questions:

Research Question 1: How does local dependence bias parameter estimates and which parameter estimates are most affected?

Research Question 2: How can model diagnostics be used to detect local dependence?

Research Question 3: How can local dependence be corrected in the three-indicator latent class model?

To answer the first question, expeculation¹ (Biemer 2011) will be used to illustrate the effects of the various types of local dependence on the parameter estimates. To address the second and third question, we examine the current literature and propose a process by which a three-indicator model can be tested for local dependence and what corrective steps can be taken if dependence is identified. We then apply our proposed process on data from the 2007 National Inmate Survey (NIS) —a nationally representative survey of prison inmates that estimated the 12-month prevalence of sexual victimization in prisons (Beck and Harrison 2007). Our analysis only considered victimizations by other inmates (inmate on inmate) although data on victimizations by staff members (staff on inmate) were also collected in the NIS.

Because one of the key objectives of the survey was to evaluate classification error in reports of prison rape and sexual assault, the NIS embedded five indicators for each type of victimization in the questionnaire. This design affords a rare opportunity to consider issues associated with three-indicator models by simulating how the estimates based upon fewer indicators compared to idealized estimates derived from the best full, five-indicator model. Ten models can be formed by taking all possible combinations of three indicators. Each of these models will be fit using an approach we developed for assessing and correcting for local dependence. To assess the validity of each model both before and after our corrections are applied, we compare our results to the ideal estimates.

Modeling Local Dependence

This section reviews the methods available for diagnosing local dependence in a latent class model and extending the standard model to appropriately account for it when it is present. It then considers each of the three causes of violation and, using an artificial population

¹Expeculation creates an input data table with frequencies that are equal to the expected cell frequencies under the assumed population model. This technique is appropriate if interest is only on the expected values or biases of the estimators rather than their standard errors or sampling distributions.

approach, demonstrates the effects of the local dependence by cause on parameter estimates, particularly the response probabilities. Methods for mitigating the effects by cause are also described.

A number of diagnostic techniques for identifying local dependence in latent class models have been proposed in the literature. Sepulveda et al. (2008) provide a good summary of these methods as well as presenting their own approach. Hagenaars (1988) proposed the “extended latent class model” obtained by adding *direct effects* (i.e., two-way indicator interaction terms) to model the dependence. Other approaches, for example, the correlation residual plot (Qu, Tan, and Kutner 1996) and the biplot (Sepulveda et al. 2008) use graphical displays to identify local dependence, while others, such as the likelihood ratio chi-square method (Espeland and Handelman 1989), comparison of standardized residuals (Hagenaars 1988), the log odds ratio check (LORC; Garrett and Zeger 2000), and bivariate residual (Magidson and Vermunt 2004), compare some version of the observed frequencies of each indicator pair with expected frequencies in which a statistically significant deviation indicates local independence. When reviewing these methods, we had two primary considerations: (a) guidance provided by the method to correctly specify the model and (b) ease of use for practical survey applications.

Regarding (a), except for the biplot, each of these methods only indicates whether a pair of variables is locally dependent. They do not indicate whether this dependence occurs in some of the latent classes or all of them (Sepulveda et al. 2008). That is, these methods indicate the need to add a direct effect (i.e., *AB*, *BC*, or *AC* interaction terms) to the model, but do not indicate when a three-way parameter involving the latent variable is necessary (e.g., an *ABX* term). Therefore, these methods would potentially identify only part of the problem.

Regarding (b), we found that the graphical approaches, such as the biplot method, were cumbersome to create, especially when several grouping variables were incorporated. The biplot method was designed to compare medical diagnoses between several physicians using only a few, if any, model covariates. However, appropriate use of grouping variables is required to address latent heterogeneity, as we discuss below. In survey applications, there may be several grouping variables and their interactions which greatly increase the complexity of biplot graphical analysis.

To gain insights regarding the biasing effects of local dependence, a technique referred to as *expeculation* (Biemer 2011) was used to assess the bias in parameter estimates under a wide range of scenarios. First, we generate an artificial population that follows a standard latent class model except that it includes a specific violation of local independence—bivocality, behaviorally correlated error, or latent heterogeneity. The standard model is then fitted to the expected cell frequencies from the population model yielding parameter estimates. The difference between an estimate from the standard model and the corresponding known population parameter is the bias in the estimate induced by the specific type of model misspecification. The results from these expeculation experiments will now be summarized.

Bivocality

Two indicators A , corresponding to the latent variable X , and C , corresponding to latent variable Y , will be univocal if and only if the correlation between X and Y is 1. Otherwise, they are bivocal. Assume that $\pi_1^X = \pi_1^Y = \pi$, say, which is made primarily to simplify the formulation and relaxing it will not alter the main findings of the numerical experiment. It seems plausible, however, and should be true, at least approximately in practice, because if the two bivocal measures are estimating similar outcomes, their prevalence probabilities should be approximately equal. With this assumption, it can be shown that

$$\rho_{xy} = \text{Corr}(X, Y) = \frac{\pi_{111}^{XY} - \pi}{1 - \pi}. \quad (2)$$

Note that if $\pi_{111}^{XY} = 1$, then X and Y are identical, $\rho_{xy} = 1$, and A and B are univocal. If $\pi_{111}^{XY} \neq 1$, then X and Y are not the same latent variable and A and B are bivocal. The severity of bivocality can be gauged by the deviation of π_{111}^{XY} from 1. Note that $\pi_{111}^{XY} = \rho_{xy}(1 - \pi) + \pi$ and $\pi_{010}^{XY} = \rho_{xy}\pi + (1 - \pi)$. Thus, the expected frequency, m_{abc} , of the cell (a, b, c) of the ABC table can be written as

$$\begin{aligned} m_{abc} &= N \sum_x \sum_y \pi_{xy}^{XY} \pi_{abc|xy}^{ABC|XY} \\ &= N \sum_x \sum_y \pi_{x|y}^{X|Y} \pi_y^Y \pi_{a|x}^{A|X} \pi_{b|x}^{B|Y} \pi_{c|y}^{C|Y}. \end{aligned} \quad (3)$$

It can be shown by substituting π for π_y^Y and applying equation (2) that the expected frequency is a function only of π , ρ_{xy} and the error probabilities $\pi_{a|x}^{A|X}$, $\pi_{b|x}^{B|X}$, and $\pi_{c|y}^{C|Y}$.

To assess the bias due to bivocality, expected cell frequencies for the ABC table were generated for values of ρ_{xy} in the range 0 to 1 at 0.1 increments. At each value of ρ_{xy} , the false-negative probabilities of each indicator set equal (i.e., $\pi_{011}^{A|X} = \pi_{011}^{B|X} = \pi_{011}^{C|Y} = \pi_{011}$, say) where π_{011} varied over the range [0.1, 0.3]. The false-positive probabilities (i.e., $\pi_{011}^{A|X} = \pi_{011}^{B|X} = \pi_{011}^{C|Y} = \pi_{011}$, say) were $\pi_{110} = 0.01$ for each indicator to simulate a sensitive event. Three values of the prevalence probability, π , were considered—.02, .05, and .1—to simulate a rare event. The bias in each parameter was computed as the difference between the parameter estimates from the standard model and assumed population model.

Figure 1 displays the percentage relative bias in the bivocal indicator by correlation level for three different false-negative rates when the prevalence is 0.02. Similar results were found when the prevalence was increased to 0.05 and 0.1. These results suggest that the bivocal indicator, C , behaves like a poor indicator of X in the analysis when Y is left out of the model. For example, the estimates of the false-negative probabilities, $\pi_{011}^{C|X}$, are biased as

estimators of the model probability $\pi_{011}^{C|Y}$. Further, the bias is inversely proportional to the correlation between X and Y or, in other words, the bias increases as the severity of bivocality increases, which is expected. Interestingly, bivocality in C did not impart a bias in the error parameters for the two univocal indicators (i.e., A and B) and, thus, the estimates for the two univocal indicators remain unbiased.

Mitigation Tactics

These execution results suggest that bivocality should be suspected for any indicator having a very high estimated error probability. In the simulations, we found that the LORC method was unreliable for diagnosing and correcting for bivocality in three-indicator models. Rather than relying on the LORC, one option is always test for a direct effect between the bivocal indicator and one or both of the other two indicators. Additional degrees of freedom to accommodate the degrees of freedom for the direct effect can be obtained if the model contains grouping variables. In fact, as we will discuss subsequently, grouping variables can be added to the model to mitigate latent heterogeneity when it is suspected and when good predictors of the latent source of heterogeneity are available. However, increasing model degrees of freedom does not guarantee identifiability and additional model constraints may be needed. For example, in the analysis of sensitive events, false positives for some indicators may be negligible (i.e., 0.001 or less). If this is true for the two univocal indicators (say, A and B), then constraining their false-positive probabilities to 0 will allow the direct effect between one or both of the univocal indicators and the bivocal indicator to be estimated (i.e., AC or BC). However, we only recommend making this constraint when identifiability cannot be obtained through the addition of grouping variables. Simulations suggest that, unless they are negligibly small, constraining the false-positive rate to 0 can lead to estimates of the false-negative probabilities that are considerably overestimated.

When there are four or more indicators, augmenting the model with a second latent variable is an option. Ideally, at least two of the indicators should be strongly associated with each latent variable. This is best accomplished by judicious study design rather than simply hoping for the availability of good indicators of both constructs at the analysis stage. In our study, treating bivocality first before moving on to address the other two causes of local dependence produced the best-fitting model.

Behaviorally Correlated Error

If A and B are dichotomous indicators for the latent variable X , then behaviorally correlated error can be expressed as the correlation between A and B given X as follows:

$$\text{Corr}(A, B|X) = \rho_{ab|x} = x \sqrt{\frac{\pi_{111}^{A|X} \pi_{111}^{B|X}}{\pi_{011}^{A|X} \pi_{011}^{B|X}}} (\lambda_1 - 1) + (1 - x) \sqrt{\frac{\pi_{110}^{A|X} \pi_{110}^{B|X}}{\pi_{010}^{A|X} \pi_{010}^{B|X}}} (\lambda_0 - 1), \quad (4)$$

where $\lambda_x = \frac{\pi_{11|x}^{B|AX}}{\pi_{11|x}^{B|X}}$ for $x = 0, 1$. Note when $\lambda_x = 1$ for $x = 0, 1$, $\rho_{ab|x} = 0$ and A and B are

locally independent. Thus, local dependence occurs if either λ_1 or λ_0 differs from unity. The expected cell frequencies for the ABC table can be written as

$$m_{abc} = N \sum_x \pi_x^X \pi_{ab|x}^{AB|X} \pi_{c|x}^{C|X}, \quad (5)$$

where

$$\begin{aligned} \pi_{11|x}^{AB|X} &= \lambda_x \tilde{\pi}_{11|x}^{AB|X}; \quad \pi_{10|x}^{AB|X} = \pi_{11|x}^{A|X} - \lambda_x \tilde{\pi}_{11|x}^{AB|X}; \quad \pi_{01|x}^{AB|X} = \pi_{11|x}^{B|X} - \lambda_x \tilde{\pi}_{11|x}^{AB|X}; \quad \pi_{00|x}^{AB|X} = \pi_{11|x}^{B|X} - \pi_{11|x}^{A|X} \\ &+ \lambda_x \tilde{\pi}_{11|x}^{AB|X}; \quad \text{and} \quad \tilde{\pi}_{11|x}^{AB|X} = \pi_x^X \pi_{11|x}^{A|X} \pi_{11|x}^{B|X} \end{aligned}$$

Analogous cell probabilities can be derived for models with any number of indicators.

To illustrate the biasing effects of behavior correlation, we consider the special case where $\lambda_1 = \lambda_0 = \lambda$ for $\lambda \neq 1$ and generated cell frequencies satisfying equation (5). Correlation in equation (4) was varied from 0 to 1 by varying λ over an appropriate range. The standard latent class model was then fit to these data and the bias induced by the behavior correlations was estimated by the difference between the standard latent class model estimates and the corresponding parameters of the population model. As we did for bivocality expectation, at each value of the correlation, the false-negative probabilities varied together in the range of $0.1 \leq \pi_{0|1} \leq 0.3$, the false-positive probabilities were fixed at .01 and these expectations were repeated for $\pi_1^X = .02, .05$, and $.1$.

The results of the expectation found that behavior correlation will bias all the parameter estimates, not just those associated with the two indicators that are conditionally correlated. Figure 2 shows that, for the prevalence probability, the magnitude of the relative bias usually increases as $\rho_{ab|x}$ increases; however, the direction of the bias depends on the magnitude of the false-negative probability. Specifically, when the false-negative probability is either small (i.e., .1) or large (i.e., .3), the magnitude of bias increases with increasing $\rho_{ab|x}$.

However, when the false-negative probability is at a moderate level (i.e., .2), the bias varies slightly around 0 regardless of $\rho_{ab|x}$. This phenomenon is caused by the canceling effects of false-positive and false-negative classifications. When the false-negative probability is small, the number of false positives overwhelms the number of false negatives causing a negative bias. When the false-negative probability is large, the opposite effect occurs. However, when the false-negative probability is at a moderate level, the numbers of false-negative and false-positive classifications are equalized and the bias is close to 0. Similar results were obtained for $\pi_1^X = 0.05$ and $\pi_1^X = 0.10$.

For the classification error rates, the level of bias depends on the magnitude of the error. For example, for the false-negative rate, the relative bias is greatest when the actual false-negative rate is smallest (i.e., 0.1). The relative bias then decreases as the actual false-negative rate increases. However, the false-positive rate, which is fixed at a negligible level,

acts in an opposite fashion, having the largest relative bias when the false-negative rate is largest (i.e., 0.3) and smallest when the false-negative rate is smallest (i.e., 0.1). Similar results were obtained for all prevalence levels examined.

Mitigation Tactics

The LORC method (Garret and Zeger 2000) can be used to identify and compensate for behavior-correlated indicators. This method identifies pairs of indicators that are locally dependent so that a direct effect for the pair can be added to correct the model. Direct effects are added one at a time starting with the pair exhibiting the highest correlation. The LORC test is repeated and, if additional correlated pairs are identified, a second direct effect is added, again only the most offensive pair, and so on until no further dependencies are identified. An optional step is to test for correlated error in the latent classes by replacing the two-way interactions with three-way interactions (i.e., XAB , XBC , or XAC) testing first for identifiability and then for significance. As previously noted, adding direct effects to three-indicator models is not possible without extending the model either by adding grouping variables—for their substantive interest or to address latent heterogeneity—or by imposing constraints, or both.

As noted above for bivocality, the LORC method is not effective in identifying locally dependent indicators in three-indicator models. Instead, we recommend testing each direct effect (AB , BC , and AC) separately in a stepwise fashion, provided an identifiable model can be obtained (e.g., through the addition of grouping variables as discussed below). This approach worked well in our empirical analysis.

Latent Heterogeneity

To demonstrate the effects of latent heterogeneity, consider the simplest case where the error probabilities vary across the categories of an *unobserved*, dichotomous grouping variable H but are homogeneous within the categories of H . In other words, the indicators are locally independent given the variable, H . To see how this induces unconditional local dependence, suppose A and B satisfy local independence given H ; that is,

$$\pi_{ab|XH}^{AB|XH} = \pi_{a|Xh}^{A|XH} \pi_{b|Xh}^{B|XH}. \quad (6)$$

In order for unconditional local independence to hold, we must have

$$\frac{\pi_{ab|x}^{AB|X}}{\pi_{a|x}^{A|X} \pi_{b|x}^{B|X}} = \sum_h \pi_h^H \frac{\pi_{a|Xh}^{A|XH} \pi_{b|Xh}^{B|XH}}{\pi_{a|x}^{A|X} \pi_{b|x}^{B|X}} = 1, \quad (7)$$

which, in general, is not true unless $\pi_{a|x}^{A|X} = \pi_{a|Xh}^{A|XH}$ and $\pi_{b|x}^{B|X} = \pi_{b|Xh}^{B|XH}$ for all h ; that is, the error probabilities are homogeneous across the levels of H .

To illustrate, consider a population with parameters given in Table 1. This set of parameters was chosen to represent a typical high-risk/low-risk scenario. The proportion of the

population in the high-risk group is given by $\pi_1^H = 0.2$ which has a high prevalence, $\pi_{111}^{X|H} = 0.8$, whereas the prevalence in the low-risk group is $\pi_{111}^{X|H} = 0.2$. We further assume $\pi_{1111}^{A|XH} = \pi_{1111}^{B|XH} = 0.7$ and $\pi_{1112}^{A|XH} = \pi_{1112}^{B|XH} = 0.9$. Finally, we assume that the false-positive probabilities are the same for both groups and quite small; namely, $\pi_{1101}^{A|XH} = \pi_{1102}^{A|XH} = \pi_{1101}^{B|XH} = \pi_{1102}^{B|XH} = 0.01$. Then $\pi_{111}^{A|X} = \pi_{111}^{B|X} = [0.2(0.7) + 0.8(0.9)] = 0.86$. Since A and B are locally independent conditional on H , we compute the ratio in equation (7) for $a = 1$, $b = 1$, $x = 1$ as $[0.2(0.7)^2 + 0.8(0.90)^2]/(0.86)^2 = 1.009$ which differs from 1 indicating *local dependence*. If H is known, adding the terms $\pi_{a|XH}^{A|XH}$ and $\pi_{b|XH}^{B|XH}$ to the model eliminates any bias due to heterogeneity.

Suppose H is unknown and let G be a dichotomous, *observed* grouping variable that is intended to serve as a proxy for H in the latent class model (i.e., G is selected so that $\pi_{i|Xh}^{I|XH} \cong \pi_{i|Xg}^{I|XG}$ when $g = h$ for all indicators, $I = A, B$, and C). For example, for crime victimizations, G may be an indicator for low-income persons. Consider the latent class model with cell probabilities

$$\pi_{gabc}^{GABC} = \sum_x \pi_x^X \pi_{g|x}^{G|X} \pi_{a|xg}^{A|XG} \pi_{b|xg}^{B|XG} \pi_{c|xg}^{C|XG}. \quad (8)$$

The ability of the grouping variable G to account for possible heterogeneity in the error probabilities induced by H depends upon the correlation between G and H . The greater the correspondence between G and H , the better the ability of G to explain variations across the categories of H and the smaller the bias will be. Thus, the biasing effects of heterogeneity due to H on the model estimates can be expressed as a function of this correlation which is given by

$$\rho_{GH} = \text{Corr}(G, H) = \frac{\pi_1^G(\pi_{111}^{H|G} - \pi_1^H)}{\sqrt{\pi_1^G(1 - \pi_1^G)\pi_1^H(1 - \pi_1^H)}}. \quad (9)$$

Suppose $\pi_1^G = \pi_1^H$ and $\pi_{111}^{G|H} = 1$ (i.e., G is essentially identical to H). Then, from equation (9), $\rho_{GH} = 1$ and the heterogeneity is fully explained by G . However, if either $\pi_1^G \neq \pi_1^H$ or $\pi_{111}^{H|G} < 1$, then G is not a perfect proxy for H , $\rho_{GH} < 1$, and the model is misspecified in that G does not fully compensate for the heterogeneity induced by H .

To fix the ideas, we assume that $\pi_1^G = \pi_1^H = \pi^*$, that is, both G and H have the same marginal probabilities, although they may still be unrelated variables. This assumption simplifies the formulae but does not alter the main findings of our expectation analysis. Instead, we focus on the bias when $\pi_{111}^{H|G} < 1$, that is, when G is not a perfect predictor of H . Then equation (9) can be rewritten as

$$\rho_{GH} = \frac{\pi_{111}^{H|G} - \pi^*}{1 - \pi^*}, \quad (10)$$

and thus, $\pi_{111}^{H|G} = (1 - \pi^*)\rho_{GH} + \pi^*$. Therefore, given π^* and ρ_{GH} , the joint distribution of H and G is fully specified.

Expeculation was used to demonstrate the biasing effects of latent heterogeneity for specified values of π^* and ρ_{GH} through the expression for the expected cell frequencies for the GABC table given by

$$m_{gabc} = N \sum_h \sum_x \pi_g^G \pi_{g|h}^{G|H} \pi_{x|h}^{X|H} \pi_{a|h}^{A|HX} \pi_{b|h}^{B|HX} \pi_{c|h}^{C|HX}. \quad (11)$$

The model in equation (8) will be fit to the frequencies given by equation (11) to evaluate the bias in using G as a proxy for H , varying correlation between H and G . We also varied π^* from 0.5 to 0.9 in our analysis with similar results to those shown here.

We use the parameter values in Table 1 to illustrate the bias associated with estimating the population false-negative probability, $\pi_{011}^{A|X}$ and the inability for G to correct of local dependence bias unless $\rho_{GH} = 1$.

A typical example of the expeculation results is shown in Figure 3 for the parameters in Table 1. In this figure, estimates of $\pi_{011}^{A|XG}$, $\pi_{012}^{A|XG}$, and $\pi_{011}^{A|X}$ are plotted as a function of the correlation between G and H . At $\rho_{GH} = 0$, the absolute values of the bias in false-negative rates are maximized. The estimates of the false-negative probabilities for $G = 1$ and $G = 2$ are the same as is estimate of the overall false-negative probability, namely, .175. As the correlation increases to 1, the group-level and overall estimates separate and smoothly converge to their true rates—that is, $\pi_{011}^{A|XG} = \pi_{011}^{A|XH} = 0.3$ and $\pi_{012}^{A|XG} = \pi_{012}^{A|XH} = 0.1$, and $\pi_{011}^{A|X} = 0.2$ —and the biases converge to 0. The prevalence rates, which are not shown in the figure, follow essentially the same pattern—a maximum bias at $\rho_{GH} = 0$ but approaching zero bias as ρ_{GH} approaches 1. The same pattern held over the ranges of values for group sizes, prevalence probabilities, and error probabilities investigated in the study.

Mitigation Tactics

Latent heterogeneity can be addressed by the addition of grouping variables, G , that can serve as proxies for one or more latent variables that induce heterogeneity. Typical grouping variable include respondent demographic variables, paradata² related to the survey process, or any other variable that is potentially of measurement error. An ideal grouping variable is a

².Paradata are data about the process by which survey data were collected (e.g., mode of interview and time of interview). See, for example, Kreuter (2013).

good predictor of the latent variable, H , that represents the source the heterogeneity in the indicator error probabilities. An additional benefit of this approach is the increase in model degrees of freedom can be used to address other sources of local dependence, for example, by incorporating direct effects, as noted above.

Because latent heterogeneity is likely in any LCA, a reasonable first step in the model selection process is to add grouping variables to the measurement component of the model guided by plausible theories regarding factors likely to be related to underreporting of sensitive behaviors or the variations in measurement error among population domains. In most situations, either the likelihood ratio test (for nested models), Bayesian Information Criterion (BIC; for nonnested models), or Wald's statistic can be used to determine the combination of grouping variables that achieve the best-fitting model (Biemer 2011) and will be most effective at reducing heterogeneity effects.

As previously noted, effective modeling of latent heterogeneity requires advanced planning as well as a good understanding of the causal factors of classification error. It may be necessary to supplement the questionnaire or paradata to obtain variables solely purposed for modeling heterogeneity. As an example, if interviewers are expected to be an important cause of heterogeneity, data on interviewer experience, performance ratings, and so on, might be useful for modeling purposes. Such data are usually not available to data analysts and often must be specially requested.

Results Summary

The expectation experiments clearly demonstrate that each cause of local dependence biases the model estimates and provides insight regarding how one can mitigate these biasing effects in a data analysis. Our results showed that a single bivocal indicator, say C of Y , will perform similar to a weak indicator of X . Despite indicator C 's bivocality, valid inference is still possible for the other univocal indicators in the model provided other sources of local dependence are not active. In simulations, mild bivocality (i.e., values of ρ_{XY} near 1) behaves somewhat like behavioral correlation and it can often be mitigated in the same way—that is, by adding direct effects. Bivocality is perhaps the most serious violation with three indicators when X and Y are only weakly correlated. Then the model will then be weakly identifiable because it approaches a condition similar to a two-indicator model. In the extreme case where X and Y are essentially unrelated, the model will be unidentifiable because then there are essentially only two indicators of X —namely, A and B —with C contributing essentially no information about X .

Indicators that are behaviorally correlated bias both the prevalence and the classification probability estimates. The magnitude of the bias increases as the correlation increases (assuming positive correlation). The addition of direct effects seems to be the appropriate strategy for dealing with this source of bias; however, that approach is not feasible in three-indicator models unless the model is extended or constraints are imposed on some parameters. For sensitive topics, constraining the false-positive probabilities to 0 may be feasible for some indicators and will often provide valid inferences. However, as noted earlier, we recommend attempting other mitigating strategies before incorporating these constraints into the model.

Latent heterogeneity primarily affects the misclassification probability estimates and, thus, steps should be taken to correctly specify the measurement component of the model to address potential variations in error probabilities across population domains. The success of this strategy depends upon the covariates available for modeling heterogeneity. If the LCA is anticipated by the research design, a richer set of covariates for data analysis can be collected and the bias due to latent heterogeneity can be greatly reduced. Otherwise, the success of this modeling strategy is uncertain.

These simulation experiments enhance our knowledge and understanding of the effects of local dependence in that they suggest which parameters are affected and to what extent. Understanding the marginal impact of each type of model violation is useful, but still limited because multiple violations are likely to occur in practice. Opposing (cancelling) as well as concordant (Cumulative) biases are possible making it difficult to provide general statements regarding how any particular model failure will influence inference.

We advocate a general approach of systematically testing and correcting for each type of local dependence separately using diagnostics and mitigation strategies that are tailored to the specific cause. Our proposed strategy begins with a review the questionnaire items that generated the indicators to look for any evidence of bivocality. When bivocality is suspected, the first step in the modeling process should be to address it by adding a second latent variable, Y , if an identifiable model can be achieved. Otherwise, modeling for bivocality can be deferred until the final step of the process.

As the second step of process, we suggest adding and testing covariates that are predictive of the variation in error probabilities to address potential latent heterogeneity. This makes sense, given that the addition of grouping variables will provide additional degrees of freedom for addressing possible behavioral correlation (and any possible residual bivocality not addressed in the first step) in the third step. Finally, the techniques described above for detecting and repairing behavioral correlation can be implemented. This three-step approach can be summarized as follows:

1. Check the construction of the indicators for *prima facie* evidence of bivocality. If bivocality is suspected, either (a) introduce a new latent variable for the bivocal indicator/indicators or (b) add a direct effect between the bivocal indicator and one or both of the two univocal indicators, testing for model improvement as appropriate.
2. Address any potential issues with latent heterogeneity by adding grouping variables predictive of the false-negative error variation in the population.
3. Test for behavioral correlation and other residual local dependencies among the indicators by adding direct effects between indicators use the LORC test as a guide.

Although this approach has worked well in our applications, the ordering of the steps is not critical. What is critical, however, is that each source of local dependence be considered and mitigated to the extent possible in the modeling process. As previously noted, it may not be possible to execute step 1 in three-indicator models before executing step 2 which has an

additional benefit of creating degrees of freedom for model enhancement. Our experience with the 2007 NIS and many other data sets suggest that step 3 should be the last step in the model-building process. The next section illustrates this three-step strategy to address local dependence while achieving model parsimony in LCA in an analysis of the 2007 NIS data.

Empirical Results

The 2007 NIS is a probability-based survey of 146 confinement prisons. Prisons were selected with unequal probabilities (probability proportional to size) and inmates within a prison were selected with simple random sampling. A total of 23,398 inmates responded to the survey yielding a response rate of 72 percent (Beck and Harrison 2007). Because of the high level of interest in evaluating classification error for this survey, five questions on past-year inmate on inmate sexual abuse were embedded in the questionnaire. Table 2 provides the wording of these questions as they appeared on the questionnaire. A majority of the sample was interviewed using audio computer-assisted interviewing (ACAS) while a random subsample of 530 respondents was interviewed using a paper and pencil interviewing questionnaire. Only respondents in the ACASI mode were asked the full set of questions in Table 2 and are included in the present analysis. Therefore, the sample size for our analysis is 22,868 respondents.

In what follows, the latent variable, X , represents a respondent's true past-year victimization status with indicators (A , B , C , D , and E) corresponding to the responses to the five questions in Table 2. The primary purpose of embedding these five indicators in the questionnaire was to use LCA to evaluate the extent of underreporting of sexual assault by inmates and compare the error properties of alternative questions for measuring inmate on inmate sexual assault. The results could aid decisions regarding the accuracy of prison rape statistics as well as what questions work best and should be used in future surveys of this type. For the present purposes, these data will be used only to demonstrate how local dependence biases these indicators and whether the bias can be effectively reduced using the three-step modeling strategy proposed in the previous section.

Because inmate on inmate sexual assault is a sensitive event, we expected the probability of false reports of abuse to be negligibly small (i.e., it is unlikely that an inmate would falsely claim to be sexually victimized). Conversely, we expect underreporting (false-negative reports) to be substantial because truly victimized inmates might indicate that they were not victimized because of the traumatic nature of the experience or fear of reprisal from the perpetrator.

As described in the second section, the primary focus of this analysis is to investigate the validity of LCA for three-indicator models of sensitive behaviors. Our proposed modeling strategy will be applied to all 10 possible three-indicator models that can be formed from the five indicators in Table 2. In each case, the effectiveness the three-step strategy to correct for local dependence will be assessed by comparing the estimates of a three-indicator model derived from the three-step approach to a gold standard set of estimates derived from the full (five-indicator) model.

A key feature of the empirical study is the creation of a set of gold standard estimates for judging the validity of the three-indicator models. Using all five indicators in Table 2 as well as a pool of 15 covariates for modeling the heterogeneity, we obtained a five-indicator latent class model that best fits the data. This model, referred to as the *best full model* and the process by which it was created will be described subsequently. Because this model has been optimized incorporating all five indicators, it will be used as a “gold standard” for judging the accuracy of the estimates produced by the three-indicator models; that is, the best full model estimates will be treated as population parameters in the empirical analysis.

Obtaining the Gold Standard Estimates

The LEM software (Vermunt 1997) was used to fit all the models in this article, ignoring the complex sample design of the NIS. Interest in this article is focused solely on the efficacy of the proposed modeling approaches and estimates of the prevalence and error rates will not be interpreted. Therefore, inferences to the 2007 NIS population should be avoided.

The best full model was developed using the three-step approach described in the previous section to systematically address all sources of local dependence (bivocality, heterogeneity, and behavioral correlation). For step 1, a substantive review of the items in Table 2 was conducted, comparing the wording of each the survey question to the definition of the latent variable of interest (i.e., “sexual assault”). From this review, it was quite obvious that indicators *A*, *B*, and *C* are univocal for sexual assault (*X*), while *D* and *E* relate to a different underlying latent variable (*Y*) since they address specific sexual acts (namely, “oral or anal sex”) rather than more general types of sexual contact. Thus, indicators *A*, *B*, *C*, *D*, and *E* are bivocal and both *X* and *Y* are required in the model to appropriately model their true values. Further, note that if an inmate was a victim of a sexual act as defined by *Y*, then he or she was subject to sexual victimization as defined by *X* with certainty (i.e., $\pi_{110}^{Y|X} = 0$).

This constraint introduces a structural zero in the *XY* table because the cell corresponding to *X* = 0 and *Y* = 1 occurs with probability 0. This model, denoted by $\{AX BX CX DY EY\}$ (using Goodman’s hierarchical model notation) is the base model and referred to as *model 0* in Table 3.

In the next step, latent heterogeneity was addressed. The analysis data set contains eight demographic variables and seven paradata variables that could theoretically explain the classification error variation in the population. Using a forward stepwise model selection approach as well as theoretical considerations, the best subset of these variables to include in the model was determined. The resultant model from this step is model 1 in Table 3 which contains three grouping variables; namely, *K* denoting the sexual orientation of the inmate (1 if heterosexual, 2 if otherwise), *M* denoting whether the inmate was victimized in another facility before arriving at the current facility (1 if *yes*, 2 if *no*), and *P* denoting whether the inmate reported all prior assaults (1 if *no*, 2 if *yes*). Model 2 in the Table 3 was the result of further reducing the model where terms having a *p* value greater than .3 were removed from the model.

In the third and final step, direct effects were added to the model to account for any remnants of local dependence indicated by the LORC procedure. This step resulted in model

3 shown in Table 3 which includes the terms AD and DE . Interestingly, these dependencies appear to arise from the bivocality of D and E relative to A , B , and C . Table 3 presents the model fit statistics for all four models and shows that the fit of model 3 is superior to the other models according to the $L^2 p$ value, the dissimilarity index (d), and the $BIC(L^2)$ statistic. We were unable to improve further on model 3 and, thus, this model produced the “gold standard estimates” (shown in Table 4) that were used to evaluate the three-indicator models. These estimates will be treated as true population parameters in the subsequent analysis which is conservative in that it ignores the variation of the gold standard estimates in the subsequent tests of significance related to the three indicator. Thus, the number of significant differences between the full model and three-indicator model estimates is likely overstated in our analysis.

As shown in Table 4, all false-positive rates are small (less than 0.003) as expected. The false-negative rates are highest for the bivocal indicators D and E . Indicators A and C have the lowest false-negative rates at 0.21 and 0.20, respectively. The estimated prevalence rate of the latent variable is 0.023 (i.e., an estimated 2.3 percent of all inmates have experienced some form of sexual assault). The parameter estimates between models 1 and 3 were not dramatically different indicating that the estimates are fairly robust to alternative model specifications considered, thus providing further confidence in the validity of the model.

Results of the Three-indicator Analysis

For each set of three indicators, for example, A , B , and C , the standard latent class model given by

$$\pi_{abc}^{ABC} = \sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}, \quad (12)$$

or $\{AX BX CX\}$ using shorthand notation, served as a starting point. The estimates from each model are shown in Table 5. As expected, false-positive probabilities are quite small for all indicators, while estimates of the false-negative probabilities are substantial and vary considerably across the 10 models for the same indicator, primarily as a result of local dependence. In what follows, we consider model specifications that correct for the apparent biases in the parameter estimates.

Table 5 provides ample evidence that the choice of indicators for an LCA can have a dramatic impact on the model estimates unless steps are taken to address model misspecification. The prevalence rate ranges from 1.5 percent (CDE model) to 2.1 percent (ABC model) and the false-negative rate for A ranges from 9.7 percent (ACD model) to 18.7 percent (ACE model), almost a twofold increase! Similar results can be noted for the other indicators. Correcting for local dependence in these models will reduce the variation in error rates within an indicator as well as result in greater consistency in the errors rates across all the five indicators.

The three-step modeling strategy was applied to each of the 10 three-indicator models. The diagnostic tools used to assess each model included the L^2 statistic, the dissimilarity index

(d), and the BIC based on the L^2 . For each three-indicator model, the first step in the modeling process is to inspect the questions in Table 2 for apparently bivocality and, if possible, correct for it by introducing a second latent variable. This review suggests that only model ABC is ostensibly univocal. The other nine models are bivocal— ABD , ABE , ACD , ACE , BCD , and BCE being bivocal with respect to X , and ADE , BDE , and CDE being bivocal with respect to Y . Our attempts to address bivocality using a second latent variable (Y) in the model were not successful because a two latent variable model was not identifiable without imposing constraints that would substantially biased the parameter estimates.

Recall from the simulations in the third section, a single bivocal indicator in a three-indicator model behaves like a poor indicator of the latent variable X (i.e., the indicator of Y will have high error rates with respect to X). The magnitude of the error rates will depend upon the correlation between X and Y . Further shown in the third section is that bivocality does not bias the other univocal indicators. Nevertheless, in step 3, we will attempt to model the local dependence induced by bivocality to reduce the bias in the prevalence estimates.

In step 2, heterogeneity was addressed by using a stepwise regression approach to determine which grouping variables, if any, among the 15 available variables to include in the model. For each model, the same three-grouping variables emerged as best and highly significant in the models—namely, K , M , and P defined above for the full data model.

Finally, in step 3, we consider whether direct effects are needed in the model to further correct for local dependence. Although local dependence due to bivocality is quite apparently for the nine models with a bivocal indicator, the LORC failed to identify local dependence in any of these nine bivocal models or the single univocal model. Instead, for the bivocal models, a direct effect was added between the bivocal indicator and the two univocal indicators and tested for significance using the conditional likelihood ratio test for nested models. The final models for each three-indicator model and their relevant fit statistics are shown in Table 6.

For all the 10 models, the fit statistics (i.e., dissimilarity index, the L^2 p values, and $BIC(L^2)$ diagnostics) were all more favorable after all necessary corrective measures were applied at each step of the three-step approach. After correcting for local dependence, the average prevalence was 2.4 percent with a range of 2.0 percent to 2.8 percent. The false-negative rates for indicators A and C were the smallest with an average of 20.8 percent (range of 13.9–28.5 percent) and 18.7 percent (range of 11.9–26.8 percent), respectively. The indicator with the highest false-negative rate was D with an average rate of 59.9 percent (range of 49.9–65.0 percent). The false-positive rates for all indicators were negligible with the largest being 0.32 percent for indicator E .

To evaluate any remaining bias in a latent class model after accounting for local dependence, a simple t -test was used to compare the three-indicator model parameter estimates to the corresponding gold standard estimates treating the gold standard estimate as a constant. Because the false-positive probabilities were quite small, those estimates were not tested. These results are summarized in Table 7. For all parameters, the hypothesis of estimate

equality from the three-indicator and gold standard models was rejected much more often for the standard LCA models than for the revised models (a ratio of 2.3 to 1).

Finally, we tested whether the absolute bias of the estimates was significantly reduced after revising the models for local dependence using the three-step approach. As shown in Table 8, the absolute bias for a model parameter was smaller for the revised model in at least two-thirds of the models in which that parameter appeared. For some parameters, (namely, $\pi_{2|1}^{B|X}$, $\pi_{2|1}^{D|X}$, and $\pi_{2|1}^{E|X}$), the revised model always produced a smaller absolute bias.

Conclusion

This article considered the problem of local dependence in an LCA with particular emphasis on the analysis of rare and sensitive events. Local dependence can be characterized as arising from three causes: bivocality, behaviorally correlated error, and latent heterogeneity. This characterization is beneficial to analysts because it decomposes the problem into components that can be more easily addressed in both the design and the analysis stages of a survey. Extending the work of Pepe and Janes (2007) and Sepulveda et al. (2008), the article demonstrates how each cause of local dependence can lead to quite different bias patterns that may affect inferences in different ways. Expectation experiments were used to illustrate the bias effects and develop mitigation tactics for reducing them in the model parameter estimates. Because the appropriate method for correcting local dependence bias depends on the cause, the article proposes a systematic approach for addressing each cause separately in the model-building process and then checking the model for other potential misspecifications.

Our proposed approach uses a theoretical framework to develop a process for systematically assessing each cause of local dependence and, if necessary, compensating for it in the modeling process. The case of a three-indicator model was specifically considered because that situation is common in survey work, particularly when measurement error research is only a secondary objective of the data collection. Ten models were formed by considering all possible subsets of three indicators from a real data set comprised of five indicators. Results from fitting the standard latent class model to these models found that there was a wide variation in the estimates of the common parameters among the models, indicating that all of the models were influenced to some degree by a failure of the assumption of local independence. The proposed modeling approach reduced the bias in all cases, often substantially, based upon a comparison to the best full, five-indicator model. However, the ability to fully compensate for local dependence with only three indicators was still limited in a few situations.

The empirical findings confirm that, at least for rare and sensitive outcomes, the proposed framework can substantially improve inference from three-indicator latent class models. Local dependence bias in the model estimates, both in terms of statistical equivalence to the gold standard estimates and the absolute biases, was substantially reduced. Bias mitigation is most successful when the three indicators are univocal models. In addition, of the three causes of local dependence, bivocality appears to be the most challenging. Therefore, an important lesson for questionnaire design is to avoid creating three indicators of a construct

where one is ostensibly bivocal. Otherwise, it would be beneficial for analysis to include one additional indicator to provide a pair of indicators for each construct X and Y . In addition, for strong identifiability, the correlation between X and Y should be high, say .8 or higher, based upon expectation results. For example, in our analysis, the correlation between rape (X) and sexual assault (Y) was sufficiently high that the bivocal model given by $\{XY AX BX CY DY\}$ was well identified.

Acknowledgments

Funding

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

Author Biographies

Marcus E. Berzofsky, DrPH, is a senior research statistician at RTI International. He specializes in the design and analysis of large, complex surveys. His research has focused on assessing measurement error in surveys in order to improve the accuracy of survey estimates.

Paul P. Biemer, PhD is a Distinguished Fellow, Statistics, at RTI International and Associate Director for Survey Research and Development for the Odum Institute at the University of North Carolina at Chapel Hill (UNC). His research has examined the relationships between survey design and survey error; statistical methods for assessing survey errors, particularly measurement errors; and methods for the analysis of survey data.

William D. Kalsbeek, PhD is Professor of Biostatistics, as well as Emeritus Director of the Survey Research Unit (recently renamed the “Carolina Survey Research Laboratory”), at the University of North Carolina at Chapel Hill. A fellow of the American Statistical Association, his professional interests include, but are not limited to the practice, teaching, and study of health survey research methods, with particular emphasis on the joint roles of costs and statistical efficacy in sample design and survey nonresponse.

References

- Alwin Duane F. 2007 *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley.
- Bailar Barbara A. 1968 “Recent Research in Reinterview Procedures.” *Journal of the American Statistical Association* 63:41–63.
- Beck Allen J. and Harrison Paige M. 2007 “Sexual Victimization in State and Federal Prisons Reported by Inmates, 2007” Bureau of Justice Special Rpt. No. NCJ 219414, U.S. Department of Justice, Office of Justice Programs. Washington, DC Retrieved December 16, 2009 (<http://www.ojp.usdoj.gov/bjs/pub/pdf/svsfpri07.pdf>).
- Biemer Paul P. 2004 “Simple Response Variance: Then and Now.” *Journal of Official Statistics* 20:417–39.
- Biemer Paul P. 2011 *Latent Class Analysis of Survey Error*. Hoboken, NJ: John Wiley.
- Biemer Paul P. and Lyberg Lars. 2003 *Introduction to Survey Quality*. Hoboken, NJ: John Wiley.
- Espeland Mark A. and Handelman Stanley L. 1989 “Using Latent Class Models to Characterize and Assess Relative Error in Discrete Measurements.” *Biometrics* 45:587–99. [PubMed: 2765639]

- Garrett Elizabeth S. and Zeger Scott L. 2000 "Latent Class Model Diagnosis." *Biometrics* 56:1055–67. [PubMed: 11129461]
- Groves Robert M. 1989 *Survey Errors and Survey Costs*. New York: John Wiley.
- Groves Robert M. and Daniel Cork. eds. 2008 *Ensuring the Quality, Credibility, and Relevance of U.S. Justice Statistics. Panel to Review the Programs of the Bureau of Justice Statistics. Committee on National Statistics and Committee on Law and Justice, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.
- Hagenaars Jacques A. 1988 "Latent Structure Models with Direct Effects between Indicators: Local Dependence Models." *Sociological Methods and Research* 16:379–405.
- Kreuter F 2013 *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: John Wiley.
- Krosnick Jon A. 1991 "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.
- Krosnick Jon A. and Alwin Duane. F. 1987 "An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement." *Public Opinion Quarterly* 51:201–19.
- Lazarsfeld Paul F. and Henry Neil W. 1968 *Latent Structure Analysis*. Boston, MA: Houghton Mifflin.
- Magidson Jay and Vermunt Jeroen K. 2004 "Latent Class Analysis" Pp. 175–98 in *Sage Handbook of Quantitative Methodology for Social Sciences*, edited by Kaplan David. Thousand Oaks, CA: Sage.
- Pepe Margaret S. and Janes Holly. 2007 "Insights into Latent Class Analysis of Diagnostic Test Performance." *Biostatistics* 8:474–84. [PubMed: 17085745]
- Qu Yingsheng, Tan Ming, and Kutner Michael H. 1996 "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests." *Biometrics* 52:797–810. [PubMed: 8805757]
- Sepulveda R, Vicente-Villardón JL, and Galindo MP 2008 "The Biplot as a Diagnostic Tool of Local Dependence in Latent Class Models. A Medical Application." *Statistics in Medicine* 27:1855–69. [PubMed: 18265437]
- Tourangeau R, Rips LJ, and Rasinski K 2000 *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- Vermunt Jeroen K. 1997 *LEM: A General Program for the Analysis of Categorical Data* Tilburg, the Netherlands: Department of Methodology and Statistics, Tilburg University Retrieved June 30, 2008 (<http://spitswww.uvt.nl/web/fsw/mto/lem/manual.pdf>).
- Wiggins LM 1973 *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*. Amsterdam, the Netherlands: Elsevier SPC.

Figure 1. Relative bias in the false-negative rate of a bivocal indicator by correlation level between the latent variables X and Y .

Figure 2. Relative bias in estimated prevalence probability due to correlated error as a function of the correlation and false-negative probability.

Figure 3.
False-negative rate in indicator by group as correlation between G and H increases.

Table 1.

Assumptions Used in Expeculation of Latent Heterogeneity.

Parameter	Assumption	
	High-Risk Group	Low-Risk Group
Group size (H)	0.20	0.80
True prevalence (X)	0.80	0.20
False-negative rate	0.30	0.10
False-positive rate	0.01	0.01

Table 2.

Definition of NIS LCA Indicators for Inmate-on-Inmate Sexual Victimization.

Indicator	Definition
A	<p>In the past 12 months or since you've been at the facility (if less than 12 months) has another inmate done the following?</p> <ul style="list-style-type: none"> • Use physical force to touch your butt, thighs, or penis in a sexual way? (male only) • Without physical force, use pressure to touch your butt, thighs, or penis in a sexual way? (male only) • Use physical force to touch your butt, thighs, breasts, or vagina in a sexual way? (female only) • Without physical force, use pressure to touch your butt, thighs, breasts, or vagina in a sexual way? (female only) • Use physical force to make you give or receive a handjob? (male only) • Without physical force, use pressure to make you give or receive a handjob? (male only) • Use physical force to make you give or receive oral sex? • Without physical force, use pressure to make you give or receive oral sex? • Use physical force to make you have vaginal sex? (female only) • Without physical force, use pressure to make you have vaginal sex? (female only) • Use physical force to make you have anal sex? • Without physical force, use pressure to make you have anal sex? • Use physical force to make you have some other type of sex not asked about? • Without physical force, use pressure to make you have some other type of sex not asked about? <p>= 1 if <i>yes</i> to any of the above = 2 if <i>no</i> to all of the above</p>
B	<p>In the past 12 months or since you've been at the facility (if less than 12 months), did another inmate use physical force, pressure you, or make you feel that you had to have any type of sex or sexual contact in?</p> <p>=1 if <i>yes</i> =2 if <i>no</i></p>
C	<p>How long has it been since another inmate in this facility used physical force, pressured you, or made you feel that you had to have any type of sex or sexual contact?</p> <p>=1 if past 12 months or since arrived at the facility (if less than 12 months) =2 if more than 12 months or never</p>
D	<p>Male: In the past 12 months or since you've been at the facility (if less than 12 months), did another inmate use physical force, pressure you, or make you feel that you had to have oral or anal sex?</p> <p>Female: In the past 12 months or since you've been at the facility (if less than 12 months), did another inmate use physical force, pressure you, or make you feel that you had to have oral, vaginal, or anal sex?</p> <p>=1 if <i>yes</i> =2 if <i>no</i></p>
E	<p>Male: How long has it been since another inmate in this facility used physical force, pressured you, or made you feel that you had to have oral or anal sex?</p> <p>Female: How long has it been since another inmate in this facility used physical force, pressured you, or made you feel that you had to have oral, vaginal, or anal sex?</p> <p>=1 if past 12 months or since arrived at the facility (if less than 12 months) =2 if more than 12 months or never</p>

Note. LCA = latent class analysis; NIS = National Inmate Survey.

Table 3.

Development of the Best Full Model.

Model ^a	<i>df</i>	<i>L</i> ²	<i>p</i> Value	Dissimilarity Index (<i>d</i>)	BIC (<i>L</i> ²)
0 { <i>XY AX BX CX DY EY</i> }	18	222.7	0	0.0057	42
1 { <i>KMP KXY MXY PXY AKX AMX APX BKX BMX BPX CKX CMX CPX DKY DMY DPY EKY EMY EPY</i> }	196	415.1	0	0.0101	−1545
2 Replaces <i>DMY</i> with <i>DM</i> , <i>EKY</i> with <i>EK</i> in model 1	198	466.6	0	0.0109	−1513
3 Adds <i>AD BD CE DE</i> to model 2	194	241.2	.012	0.0068	−1699

Note. BIC = Bayesian Information Criterion.

^aThe models in this table include a structural 0 via the constraint $\pi_{1|0}^{Y|X} = 0$.

Table 4.

Key Parameter Estimates From the Best Full Model.

Prevalence (%)	False-Negative Rates (%)					False-Positive Rates (%)				
π_1^X	$\pi_{0 1}^{A X}$	$\pi_{0 1}^{B X}$	$\pi_{0 1}^{C X}$	$\pi_{0 1}^{D X}$	$\pi_{0 1}^{E X}$	$\pi_{1 0}^{A X}$	$\pi_{1 0}^{B X}$	$\pi_{1 0}^{C X}$	$\pi_{1 0}^{D X}$	$\pi_{1 0}^{E X}$
2.3	20.7	38.5	20.1	60.4	41.7	0.3	0.1	0.2	0.3	0.3

Table 5.

Parameter Estimates for the Standard Three-Indicator Latent Class Model.

Model		Prevalence (%)	False-Negative Rates (%)					False-Positive Rates (%)				
		π_1^X	$\pi_{011}^{A X}$	$\pi_{011}^{B X}$	$\pi_{011}^{C X}$	$\pi_{011}^{D X}$	$\pi_{011}^{E X}$	$\pi_{110}^{A X}$	$\pi_{110}^{B X}$	$\pi_{110}^{C X}$	$\pi_{110}^{D X}$	$\pi_{110}^{E X}$
1	ABC	2.1	16.8	34.3	17.2			0.3	0.1	0.3		
2	ABD	1.6	9.4	18.4		39.8		0.7	0.2		0.2	
3	ABE	1.8	13.0	26.4			27.3	0.5	0.2			0.3
4	ACD	2.0	9.8		17.9	52.0		0.3		0.5	0.2	
5	ACE	2.0	18.8		7.8		27.8	0.5		0.3		0.2
6	ADE	1.6	16.2			36.7	15.3	0.7			0.1	0.3
7	BCD	1.7		16.7	15.9	43.9			0.1	0.7	0.2	
8	BCE	1.7		29.2	5.1		20.3		0.3	0.5		0.2
9	BDE	1.5		23.5		31.7	14.1		0.4		0.2	0.4
10	CDE	1.5			12.0	41.4	2.1			0.8	0.3	0.1

Table 6.

Summary of the Best Three-Indicator Models Obtained by the Three-Step Approach.

Indicators	Best Three-Indicator Model	<i>df</i>	<i>p</i> Value	<i>d</i>
ABC	{KMP AKX AMX APX BKX BMX BPX CKX CMX CPX}	28	0	.0057
ABD	{KMP AK AMX APX BKX BMX BP DKX DM DPX AD BD}	29	.0024	.0045
ABE	{KMP AKX AMX APX BKX BMX BPX EKX EM EPX BE}	28	.0001	.0055
ACD	{KMP AKX CKX CMX CPX DKX DM DPX AD CD}	31	.0018	.0046
ACE	{KMP AKX AMX APX CKX CMX EKX EM EPX AE CE}	29	.0014	.0049
ADE	{KMP AKX AMX APX DKX DMX DPX EKX EPX DE}	29	.0004	.0045
BCD	{KMP BKX BM BP CKX CMX CPX DKX DMX DPX BD}	29	0	.0045
BCE	{KMP BKX BMX BPX CKX CMX CPX EK EM EPX BE CE}	28	0	.0051
BDE	{KMP BKX BMX BPX DKX DMX DPX EK EM EPX DE}	29	.0001	.0047
CDE	{KMP CKX CMX CPX DKX DMX DPX EKX EMX EP DE}	28	.004	.0042

Table 7.

Statistically Significant Differences Among the Estimates of Three-Indicator Model and Best Full Model.

Parameter	Number of Models With Term	Percentage of Three-Indicator Models Having Statistically Different Estimates Compared With Best Full Model	
		Standard LC Model	“Best” LC Model
π_1^X	10	90.0	40.0
$\pi_{2 1}^{A X}$	6	66.7	50.0
$\pi_{2 1}^{B X}$	6	83.3	50.0
$\pi_{2 1}^{C X}$	6	66.7	50.0
$\pi_{2 1}^{D X}$	6	100	16.7
$\pi_{2 1}^{E X}$	6	100	16.7

Note. LC = latent class.

Table 8.

Reduction in Absolute Bias in the Estimates of the Best Versus Standard Three-Indicator Models.

Parameter	Number of Models With Term	Percentage of Models With Smaller Absolute Bias: Best Versus Standard Three-Indicator Models
π_1^X	10	80.0
$\pi_{2 1}^{A X}$	6	83.3
$\pi_{2 1}^{B X}$	6	100
$\pi_{2 1}^{C X}$	6	67.7
$\pi_{2 1}^{D X}$	6	100
$\pi_{2 1}^{E X}$	6	100