

University of Alaska Anchorage
Justice Center

Working Papers



Anchorage Community Survey 2007 Survey Sampling Design: Power and Sample Size

by

Shel Lee Evans

Justice Center
University of Alaska Anchorage

Working Paper Number 4

December 2006



<http://justice.uaa.alaska.edu/workingpapers/>
ISSN 1555-3445

Justice Center ♦ University of Alaska Anchorage ♦ 3211 Providence Drive ♦ Anchorage, AK 99508
Series coordinator: Dr. Darryl Wood ♦ ayjust@uaa.alaska.edu

Abstract:

This working paper documents the power analysis, literature review, and precision considerations contemplated in designing the Anchorage Community Survey's (ACS) 2007 sampling design. The ACS will obtain at least 30 completed surveys from individuals in each of the 55 census tracts that make up the Anchorage Municipality, allowing us to discern a fairly small effect size of 0.30 with our smallest anticipated intraclass correlation and a moderate effect size of 0.40 with our largest anticipated intraclass correlation, both at 0.80 power level. This cluster sample size and number of clusters should yield sufficient precision to allow good estimation of variance components and standard errors, acceptable reliability estimates, and reasonable aggregated measures of constructed neighborhood variables from individual survey item responses.

Introduction

This working paper documents the power analysis and considerations contemplated in designing the Anchorage Community Survey's (ACS) 2007 sampling design. It is important that the data collected have sufficient power to reject null hypotheses on a number of different topics while allowing for multilevel analyses appropriate to the nested data we will obtain (subjects nested within communities).

Toward this end, a review of the literature and power analyses using Spybrook, Raudenbush, Liu, and Congdon's *Optimal Design Software for Longitudinal and Multilevel Research, Version 1.76* were conducted. Although this software assumes balanced designs, we can expect our neighborhood samples to have unequal numbers of respondents. But with application of the Tailored Design Method (Dillman 2007), these neighborhood sample sizes should be just slightly unbalanced rather than severely unbalanced, yielding only slightly less power than completely balanced designs. Nonetheless, predicted power lines for slightly unbalanced designs remain very close to those predicted for balanced designs (Browne 2006).

Power, desired precision, cost considerations, and theoretical/practical relevance drive the size of samples within clusters and the number of clusters included in a study. The power to detect significant difference from the null hypothesis depends on the cluster size (n), the number of clusters (J), the intraclass correlation (ρ), and the effect size (δ). Each is considered in turn below.

Number of clusters (J) = 55

The number of clusters in the sample has the single greatest effect on power estimates. Of course, clusters are more expensive to add to a survey design than are individuals, since every cluster will need to contain sufficient respondents within it for accurate estimation of variance components and standard errors (Maas & Hox 2004).

There are practical and theoretical reasons to set our number of clusters at either 55 (the number of census tracts in the Anchorage Municipality) or 37 (the number of community councils in the same area). Census tracts have long been used as a neighborhood proxy in other community studies and facilitate easy linkage of our collected data to census data on socioeconomic and demographic variables. Community councils are voluntary organizations recognized by the Anchorage Municipality as advisory groups representing natural communities with common interests and a distinct identity.¹ Consequently, community councils may be a meaningful avenue for community efficacy. Because they are locally constructed on the criteria of common interests and identity, with sensitivity to natural and artificial barriers between areas and recognition of community desires about boundary-lines, community councils may also represent a measure of community that better approximates residents'

¹ For a description of the role of community councils, please see the Anchorage Municipal Code governing community councils at <http://www.communitycouncils.org/download/840.pdf> (accessed December 4, 2006).

conception of the boundaries of their neighborhood.² But because a greater number of clusters yields higher power, our design is constructed around census tracts, though we remain sensitive to the sample's potential for performing analyses both ways with reasonable power.

Cluster Size (n) = 30

The optimal n is about 30, considering the stated desire for flexibility in analyzing either census tracts or community councils. At $n=18$, we can discern effects as small as .30 at $\rho = .10$ for census tracts. At $n=30$, we can discern moderate effects for community councils at $\rho = .15$ and any improvement in that number is not rewarded with significantly better power.

Cluster sizes of 30 in each of our 55 census tracts should yield a completed Anchorage sample size of 1,650. ACS will plan for a contacted sample size of 3,300 for a conservative 50% response rate with four unique contacts using the Tailored Design Method.

Intraclass correlation (ρ) = .10 to .20³

The intraclass correlation is an estimate of the degree of association between the independent variable and the dependent variable in the population for a random effects model.⁴ Intraclass correlation “measures the extent to which individuals within the same group are more similar to each other than they are to individuals in different groups” (Dickinson & Basu 2005). In other words, the intraclass correlation reveals the degree of dependence the observations within each cluster have on one another (Raudenbush & Bryk 2002). In our design, the intraclass correlation statistic measures the proportion of variance in the outcome that exists between neighborhoods (as opposed to within neighborhoods).⁵

Spybrook, Raudenbush, Liu, and Congdon (2006) indicate that the interclass correlation for neighborhood research on mental health will be 0.05 or smaller and that the value for school achievement typically ranges between 0.05 and 0.15. But the important point to remember is that even with relatively small variance at the aggregate level (as the above ρ would indicate), we may still see moderate to large neighborhood effects (Reisig & Cancino 2004; Reisig & Parks 2000; Sampson & Jégnum Bartusch 1998; Taylor 1997; Duncan & Raudenbush 1999).

Crime Measures

Using an artificial and somewhat confounded means of clustering neighborhoods on SES and related characteristics (even when they do not share geographic boundaries), Simons, Simons, Conger, and Brody (2004) concluded that 16.6% of variation in youth conduct problems

² See Coulton, Korbin, Chan, & Su's (2001) methodological note about the divergence between researchers' operationalization of neighborhood and the ways that respondents draw the boundaries around their own perceived neighborhoods. They found that many adults drew areas as large as the census tracts that many researchers use, but the boundaries of that area differed.

³ Note that the smaller the ρ -value, the easier it is to achieve sufficient power. Estimations based on these higher values (even though only one article reports such a high intraclass correlation and does not explain its anomalous appearance) are therefore conservative.

⁴ With a fixed effects model, the estimated ρ is a *conditional intraclass correlation* that measures the degree of dependence among observations within neighborhoods that have the same level-2 predictor value (Raudenbush & Bryk 2002). But it is the ρ -value in the simpler, random effects model that we consider in calculations of the study's power to correctly reject the null hypothesis.

⁵ Note that the intraclass correlation is not the same thing as the *variance-explained* statistic. The variance-explained statistic provides the percentage of variance present at level-2 (the neighborhood level) that is accounted for by the inclusion of the variables in our model. This means that even with an intraclass correlation of .12, when we run a particular hierarchical linear model, we could see 73% of the total variance explained between neighborhoods and only 1% explained within neighborhoods (e.g., Sampson, Morenoff, & Earls 1999).

were between communities (amounting to about a 0.166 intraclass correlation). They note that this is higher than found in other studies of neighborhood effects (perhaps due to their method of artificial aggregation— see page 5).

None of the reviewed articles that used crime or violence as dependent variables reported intraclass correlation values for them.

Incivility (AKA Neighborhood Disorder)

Reisig and Cancino (2004) report an intraclass correlation of .13 for perceived incivility. Reisig and Parks (2004), using different data, report an intraclass correlation of .14 for perceived incivility.

Collective Efficacy or its Components

Sampson, Morenoff, and Earls (1999) report intraclass correlations between .10 and .13 for scales measuring collective efficacy for children. Sampson, Raudenbush, and Earls (1997) report an intraclass correlation of .21 for collective efficacy. Silver & Miller (2004) report an intraclass correlation of .12 for informal social control.

Effect size (δ) = .30 or greater

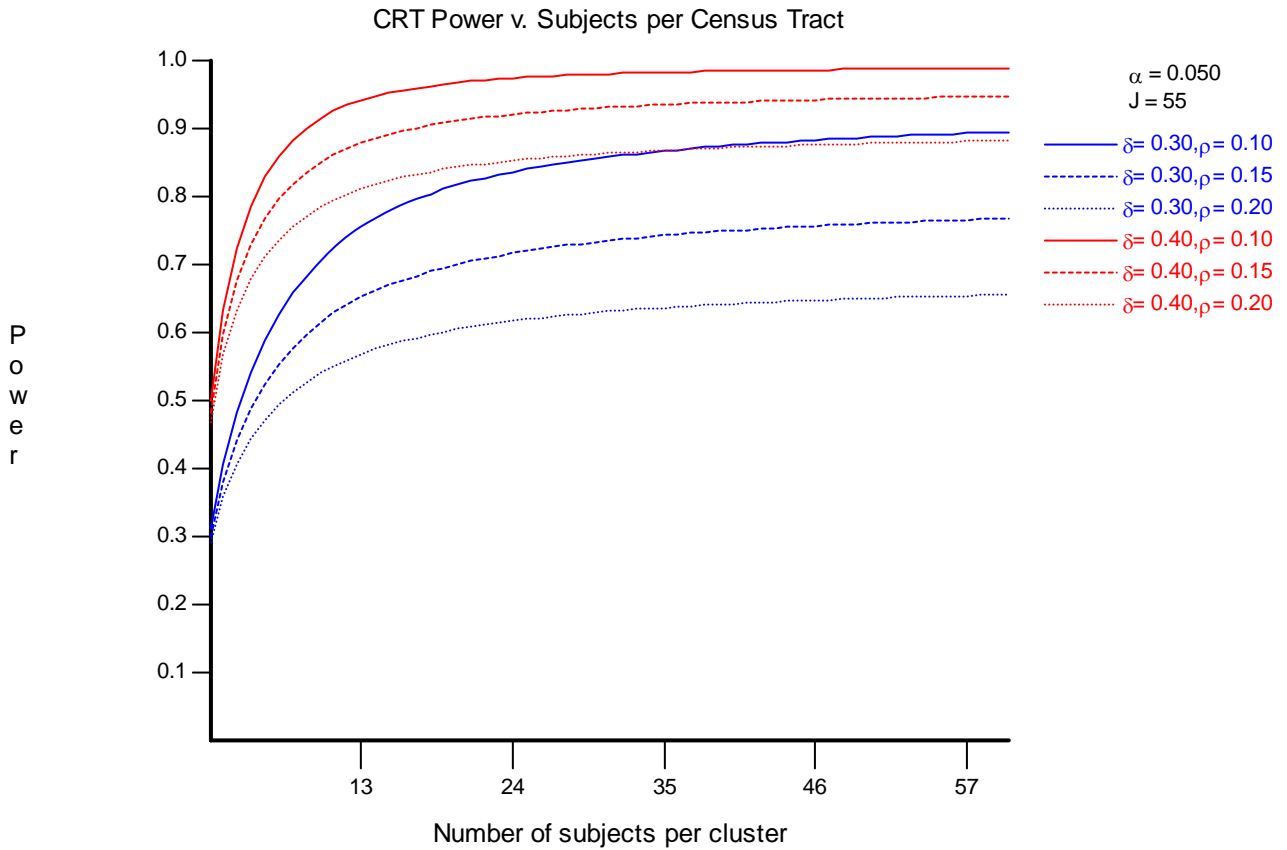
The effect size is a standardized measure of the difference between the true parameters in the population of interest and those hypothesized under the null hypothesis. The raw difference between the true and null-hypothesized parameters is divided by the degree of variability (standard deviation) in the measure in order to produce a standard effect size. The nature of the effect size varies from one statistical procedure to the next,⁶ but for our purposes, it represents a correlation coefficient at any level in our hierarchical regressions. Specified here, it indicates the magnitude of correlation we want our analyses to be capable of discerning between a specified dependent variable and any number of independent variables.⁷

Effect sizes as small as .20 to .30 are “often considered worth detecting” (Spybrook, Raudenbush, Liu, and Congdon 2006). A small effect size of .30 translates to 30% of a standard deviation unit; a large effect size of .80 is 80% of a standard deviation unit. Because .20 is the low end of the small effects, we have determined that .30 is an appropriate level to allow sufficient ease of computation across numbers of clusters (55 or 37) while capturing most effects. Use of the lower registers, in some cases, would have obviated any potential for reaching the sufficient level for power (0.80). Because smaller effects are substantively less important, are more likely to be transient, and may get lost in the error-noise of the messier conditions social researchers encounter outside of the laboratory, we can be confident in setting our effect size threshold at .30 that we will not miss meaningful correlations in our data.

⁶ For example, in clinical trials, it can represent a difference in cure rates or a reduction in symptoms among those assigned to the treatment group and those assigned to the control group. Sometimes, it just represents a standardized mean difference between groups. But ours is not an experimental design. Rather, it is a survey of the general population where no one has been assigned to any dichotomous condition category. In these cases, it is a correlation coefficient.

⁷ Most of the Optimal Design manual is written in terms of an experimental design, rather than an observational study. For an instructive presentation of power, effect size, and multilevel modeling considerations couched in observational terms, see William Browne’s (2006) presentation available at <http://www.ccsr.ac.uk/methods/festival/programme/mlm1/browne.ppt#256.1,Sample%20Size%20calculations%20in%20multilevel%20modelling> (accessed November 29, 2006).

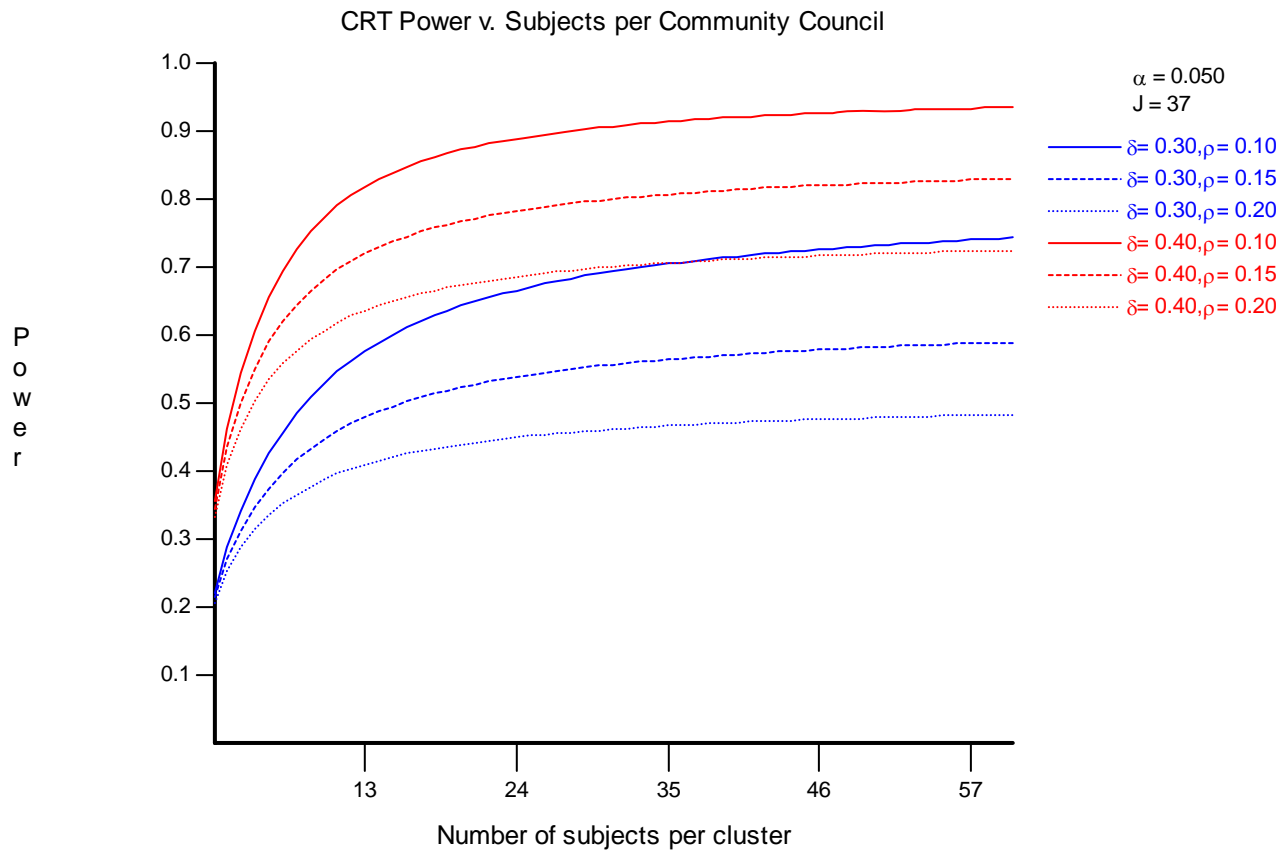
Power Graphs



The above graph shows that to reach power=0.80 and discern a fairly small effect size (.30 in blue) with our 55 census tracts, we need $n=18$ if $\rho = 0.10$ and it will never do the same if $\rho > 0.10$.

To reach power=0.80 and discern a moderate effect size (.40 in red) with our 55 census tracts, we need $n=5$ if $\rho = 0.10$, $n=7$ if $\rho = 0.15$, and $n=12$ to do the same if $\rho = 0.20$. To reach power=0.90, we need $n=9$ if $\rho = 0.10$, $n=16$ if $\rho = 0.15$, and it will never do the same if $\rho > 0.15$.

The optimal n under these modified, more conservative, and more grounded estimates for our 55 census tracts is still about 20, because at 18, we can discern effects as small as .30 at $\rho = .10$, we have ample ability to discern moderate effects, and any increase in sample size would not yield improved power. (Depictions of .20 were omitted from this estimate as it could not overcome a .60 power barrier even pushing an unfeasibly large sample size.)



The above graph shows that we cannot reach power=0.80 while discerning a fairly small effect size (.30 in blue) with our 37 community councils.

To reach power=0.80 and discern a moderate effect size (.40 in red) with our 37 community councils, we need $n=11$ if $\rho = 0.10$, $n=30$ if $\rho = 0.15$, and it will never do the same if $\rho > 0.15$. To reach power=0.90, we need $n= 28$ if $\rho = 0.10$, and it will never do the same if $\rho > 0.10$.

The optimal n for our 37 community councils is about 30, because with our smaller number of clusters as community councils, we cannot discern small effects at all, and moderate effect sizes are discernable up to $\rho =.15$ with any improvement in that number going unrewarded in terms of power.

Support from the Literature

In addition to our completed calculations justifying use of this reduced⁸ within-neighborhood sample size, the literature provides ample examples of respected scholarship doing the same when employing hierarchical analysis techniques:

- In Sampson, Raudenbush, and Earls' (1997) *Science* article, they noted that neighborhoods could reliably be distinguished at 0.80 with a sample size of 20 and at 0.91 with a sample size of 50.
- In Morenoff, Sampson, and Raudenbush's (2001) *Criminology* piece, they rely on 25 respondents per neighborhood cluster (averaging about 8,000 people in each) to achieve 0.85 reliability for their principle variable of interest: collective efficacy.
- Sampson and Raudenbush's (1999) *American Journal of Sociology* article relied on an average of 20 respondents per census tract, breaking with their usual habit of considering neighborhood cluster as their ecological unit, resulting in what they deemed an acceptable 0.68 reliability for collective efficacy.
- In 2004, the same authors published in *Social Psychology Quarterly* with 30 observations (SSO) per census block group.
- Sampson, Morenoff, and Gannon-Rowley (2002) provide "a rule-of-thumb goal" of 25 respondents within each neighborhood, based on the demonstration in Raudenbush and Sampson (1999) above.
- Rountree and Land's (1996) *Social Forces* article sampled an average of 18 respondents per neighborhood.
- Mok & Flynn (1998:413) argue that one needs only an average of 10 observations from 10 or more clusters to attain reasonably stable parameter estimates with reasonably small bias. This appears, however, to be a minority view. Most researchers expect to need more observations *and* more groups.
- Simon, Simon, Conger, and Brody (2004) artificially produced 31 clusters from block groups that were not necessarily geographically contiguous resulting in sample sizes of 7 to 56 respondents, with most "neighborhoods" containing 15 to 30 respondents.
- Coulton, Korbin, and Su (1999) published in *Child Abuse and Neglect* with 20 respondents per block group.
- Steptoe & Feldman (2001) published in the *Annals of Behavioral Medicine* with 18 respondents per postal sector.
- Perkins & Taylor's (1996) *American Journal of Community Psychology* article had an average of only 8 respondents per neighborhood.
- Elliott (DS), Wilson, Huizinga, Sampson, Elliott (A), and Rankin (1996) published in the *Journal of Research in Crime and Delinquency* with a two-location study of about 26 respondents per block group in Denver and 15 respondents per census tract in Chicago.

⁸ Previous ACIP survey designs have aimed for completed cluster sample sizes of 50 surveys in each census tract.

- Silver & Miller (2004) took their sample of 21 respondents per neighborhood cluster (average—samples ranged from 6 to 59) to *Criminology*.
- Reisig & Cancino (2004) published in the *Journal of Criminal Justice* with 16 to 72 respondents in each residential area; a mean of 36 and a median of 35 across their 31 clusters.

A Note on Precision

Power allows us to correctly reject the null hypothesis when we ought to; precision gives us reasonable expectations that we would have found these results in more than just our chance-sampling of the population. Our completed survey sample size of 1,650 (a conservative 50% response rate) will give us a 95% confidence that our observed values reflect Anchorage's true population values within ± 2.5 percentage points. But with multilevel analyses, we have an interest in precision that goes beyond these broad population comparisons as well.

The cluster sample size defined here ($n=30$) and the number of clusters (55 census tracts or 37 community councils) contemplated in the survey design outlined above do not *maximize* the available power in this study. Instead, these settings aim to produce optimal power while maintaining the precision necessary to be assured that we are not introducing additional error into the estimation of parameters in our multilevel models during analyses.

Many of our variables of interest will be constructed from individual responses to survey items that will be aggregated across neighborhoods to form neighborhood-level indicators. We must have sufficient respondents within each cluster to reduce sampling error and allow for reasonably precise estimates of these measures.⁹ Many researchers who have explicitly contemplated the issue of optimal cluster size have concluded that precise estimation of parameters and specification of models that allow for the random slopes that data sometimes demonstrate are possible only if within-cluster samples contain at least 20 to 30 individuals (Hox 1998; Kreft & de Leeuw 1998; Sampson, Morenoff, & Gannon-Rowley 2002; Snijders & Bosker 1999). Smaller cluster sample sizes are likely to underestimate standard errors and variance components (Verbeek 2000), resulting in reduced reliability and exaggerated shrinkage of the Bayes estimators (Raudenbush & Bryk 2002).

The design advanced here balances the need for sufficient power with the requirements of necessary precision.

⁹ We thank Sastry, Ghosh-Dastidar, Adams, and Pebley (2003) for articulation of this point in their own working paper outlining collection of multilevel community data.

References

- Browne, William. 2006. "Sample Size Calculations in Multilevel Modelling." Available for download at <http://www.ccsr.ac.uk/methods/festival/programme/mlm1/browne.ppt#256.1,Sample%20Size%20calculations%20in%20multilevel%20modelling>, accessed November 29, 2006.
- Coulton, Claudia J.; Jill Korbin; Marilyn Su. 1999. "Neighborhoods and Child Maltreatment: A Multilevel Study," *Child Abuse and Neglect*, 23(11): 1019-1040.
- Coulton, Claudia J.; Jill Korbin; Tsui Chan; Marilyn Su. 2001. "Mapping Residents' Perceptions of Neighborhood Boundaries: A Methodological Note," *American Journal of Community Psychology*, 29(2): 371-383,
- Dickinson, L. Miriam; Anirban Basu. 2005. "Multilevel Modeling and Practice-Based Research," *Annals of Family Medicine*, 3:S52-S60.
- Dillman, Don A. 2007. *Mail and Internet Surveys: The Tailored Design Method, 2nd Edition*. Hoboken, NJ: John Wiley and Sons.
- Duncan, Greg J.; Stephen W. Raudenbush. 1999. "Assessing the Effects of Context in Studies of Child and Youth Development," *Educational Psychology*, 34:29-41.
- Elliott, Delbert S.; William Julius Wilson; David Huizinga; Robert J. Sampson; Amanda Elliott; Bruce Rankin. 1996. "The Effects of Neighborhood Disadvantage on Adolescent Development," *Journal of Research in Crime and Delinquency*, 33(4): 389-426.
- Hox, Joop. 1998. "Multilevel modeling: When and how," in I. Balderjahn, R. Mathar, and M. Schader (eds.), *Classification, data analysts, and data highways*. New York: Springer Verlag.
- Kreft, Ita; Jan de Leeuw. 1998. *Introducing multilevel modeling*. Thousand Oaks: Sage.
- Maas, Cora J. M.; Joop J. Hox. 2004. "Robustness Issues in Multilevel Regression Analysis," *Statistica Neerlandica*, 58(2): 127-137.
- Mok, Magdalena; Marcellin Flynn. 1998. "Effect of Catholic School Culture of Students' Achievement in Higher School Certificate Examinations: A Multilevel Path Analysis," *Educational Psychology*, 18, 409- 432.
- Morenoff, Jeffrey D.; Robert J. Sampson; Stephen W. Raudenbush. 2001. "Neighborhood Inequality, Collective Efficacy, and the Spatial Dynamics of Urban Violence," *Criminology*, 39(3): 517-559.
- Perkins, Douglas D.; Ralph B. Taylor. 1996. "Ecological Assessments of Community Disorder: Their Relationship to Fear of Crime and Theoretical Implications," *American Journal of Community Psychology*, 24: 63-107.
- Raudenbush, Stephen W.; Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edition*. Thousand Oaks, CA: Sage.
- Reisig, Michael D.; Jeffrey Michael Cancino. 2004. "Incivilities in Nonmetropolitan Communities: The Effects of Structural Constraints, Social Conditions, and Crime," *Journal of Criminal Justice*, 32: 15-29.
- Reisig, Michael D.; Roger B. Parks. 2004. "Can Community Policing Help the Truly Disadvantaged?" *Crime and Delinquency*, 50(2): 139-167.

- Rountree, Pamela Wilcox; Kenneth C. Land. 1996. "Perceived Risk versus Fear of Crime: Empirical Evidence of Conceptually Distinct Reactions in Survey Data," *Social Forces*, 74(4): 1353-1376.
- Sampson, Robert J.; Dawn Jeglum Bartusch. 1998. "Legal Cynicism and (Subcultural?) Tolerance of Deviance: The Neighborhood Context of Racial Differences," *Law and Society Review*, 32(4): 777-804.
- Sampson, Robert J.; Jeffrey D. Morenoff; Felton Earls. 1999. "Beyond Social Capital: Spatial Dynamics of Collective Efficacy for Children," *American Sociological Review*, 64: 633-660.
- Sampson, Robert J.; Jeffrey D. Morenoff; Thomas Gannon-Rowley. 2002. "Assessing 'Neighborhood Effects': Social Processes and New Directions in Research," *Annual Review of Sociology*, 28: 443-478.
- Sampson, Robert J.; Stephen W. Raudenbush. 1999. "Systematic Social Observation of Public Spaces: A New Look at Disorder in Urban Neighborhoods," *The American Journal of Sociology*, 105(3): 603-651.
- Sampson, Robert J.; Stephen W. Raudenbush. 2004. "Seeing Disorder: Neighborhood Stigma and the Social Construction of 'Broken Windows,'" *Social Psychology Quarterly*, 67(4): 319-342.
- Sampson, Robert J.; Stephen W. Raudenbush; Felton Earls. 1997. "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy," *Science*, 277: 918-924.
- Sastry, Narayan; Bonnie Ghosh-Dastidar; John Adams; Anne Pebley. 2003. "The Design of a Multilevel Survey of Children, Families, and Communities: The Los Angeles Family and Neighborhood Survey," Working Paper in the Office of Population Research Princeton University Working Paper Series.
- Silver, Eric; Lisa L. Miller. 2004. "Sources of Informal Social Control in Chicago Neighborhoods," *Criminology*, 42(3): 551-583.
- Simons, Leslie Gordon; Ronald L. Simons; Rand D. Conger; Gene H. Brody. 2004. "Collective Socialization and Child Conduct Problems: A Multilevel Analysis with an African American Sample," *Youth and Society*, 35(3): 267-292.
- Snijders, Tom A. B.; Roel J. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks: Sage.
- Spybrook, Jessaca; Stephen W. Raudenbush; Xiao-feng Liu; Richard Congdon. 2006. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software*. Available for download at http://sitemaker.umich.edu/group-based/optimal_design_software, accessed November 13, 2006.
- Steptoe, Andrew; Pamela J. Feldman. 2001. "Neighborhood Problems as Sources of Chronic Stress: Development of a Measure of Neighborhood Problems, and Associations with Socioeconomic Status and Health," *Annals of Behavioral Medicine*, 23(3): 177-185.
- Taylor, Ralph B. 1997. "Social Order and Disorder of Street Blocks and Neighborhoods: Ecology, Microecology, and the Systemic Model of Social Disorganization," *Journal of Research in Crime and Delinquency*, 34: 113-155.
- Verbeek, Marno. 2000. *A Guide to Modern Econometrics*. New York: John Wiley and Sons.