Data Shop

Data Shop, a department of Cityscape, *presents short articles or notes on the uses of data in housing and urban research. Through this department, the Office of Policy Development and Research introduces readers to new and overlooked data sources and to improved techniques in using well-known data. The emphasis is on sources and methods that analysts can use in their own work. Researchers often run into knotty data problems involving data interpretation or manipulation that must be solved before a project can proceed, but they seldom get to focus in detail on the solutions to such problems. If you have an idea for an applied, data-centric note of no more than 3,000 words, please send a one-paragraph abstract to david.a.vandenbroucke@hud.gov for consideration.*

Developing a Proxy for Identifying Family Developments in HUD's LIHTC Data: Using Information on the Distribution of Units by Size

Rachel M.B. Atkins
The New School

Katherine M. O'Regan
New York University[1]

Abstract

The only existing national database on projects in the Low-Income Housing Tax Credit (LIHTC) Program has limited data on which developments serve families, a population of considerable interest to policymakers and researchers. To fill this gap, we use existing data on the size distribution of units in LIHTC projects to develop a proxy for family developments. We supplement this work with data on occupants of LIHTC developments in six states to test how well this proxy works. We estimate that this proxy would capture 92 to 96 percent of units in family developments.

[1] This article was written before the author became the Assistant Secretary for Policy Development and Research at the U.S. Department of Housing and Urban Development.

Introduction

Assessments of housing programs frequently distinguish how well such programs serve families (Khadurri, Buron, and Claminco, 2006; Khadurri, Buron, and Lam, 2004; Newman and Schnare, 1997). Although not always stated explicitly, a focus on families and their environments might arise out of heightened concern for the children they may contain or out of recognition that issues related to working-age adults may be of particular interest in housing programs. (Housing programs generally apply a loose definition of *family*, encompassing any household composition that operates as a unit, further distinguishing families from elderly families or populations requiring special services. For our purposes, we take *family* to mean a multiple-person household operating as one unit, which may or may not contain children and which would not be classified as an elderly household.) Assessing the largest federal supply-side program (the Low-Income Housing Tax Credit [LIHTC] Program) is hampered by our limited ability to identify which LIHTC developments serve (or house) families. No national data currently exist on tenants of LIHTC housing.[2] The one existing national database on LIHTC projects includes some information on whether states report that a development "targets" specific populations, including families, but those data are fairly incomplete, even among newer projects.[3] In addition, states vary on whether family is a "targeted population" in their allocation process, or if families are generally served in developments that do not target other specific groups, such as the elderly[4] or those with special needs.

In the absence of good national data on which developments serve families (whether targeted or as a remainder category), researchers have either collected the data needed for a particular state (Kawitzky et al., 2013;[5] Pfeiffer, 2009) or used proxies, such as units with at least two bedrooms (Ellen and Horn, 2012; Khadurri, Buron, and Claminco, 2006). This second method focuses on units rather than family developments as a whole, which may be more appropriate for some policy questions than others. This article develops and tests a method for identifying family developments within the national LIHTC stock, using publicly available data. We first develop this categorization scheme using the U.S. Department of Housing and Urban Development's (HUD's) LIHTC data. We then assess its performance through a combination of HUD's LIHTC data on projects and data we have collected on LIHTC tenants in six states.

HUD LIHTC Data and Methodology

This section first describes the data and relevant variables used for identifying family developments in the data. We then outline our methodological approach, which relies on observable differences in the size distribution of units in family versus nonfamily developments. The section concludes with both brief and detailed descriptions of the algorithm itself.

[2] Since 2009, states have been required to submit data on tenant characteristics to HUD, but such data are not yet available publicly.

[3] State allocation plans and the HUD LIHTC database use the term *target population* for categories declared during the allocation process. Throughout this article, we use the term *target* to indicate explicit categorization by the states, and we use the term *serve* as a broader category of developments likely to house families.

[4] The LIHTC database uses the term *elderly,* so we use that term throughout, although many state HFAs use the term *senior.*

[5] Kawitzky et al. were able to gather such data for only about one-half of their sample.

Data

We rely on two sources of data. The first is HUD's LIHTC database, which contains project-level data about developments placed in service through 2010. The HUD data contain two types of variables on populations targeted by the development: (1) whether a project targets at all (a binary variable) and (2) a series of (binary) variables for specific groups targeted, including families and the elderly.[6] Of the 36,364 developments contained in the HUD database, most (60 percent) provide either no information on targeting or indicate the development does not target, a category that may disproportionately contain developments that do serve families.[7] Even among projects placed in service in 2003 or later, when data on target population were collected more systematically, 30 percent of developments lack information on population targeted. We are interested in developing a methodology for identifying family developments within this group, those for which insufficient information is available to determine the population actually served in national data.

We supplement these project-level data with tenant-level data collected from six state housing finance agencies.[8] These data include information about the age of tenants.[9] We use these data as an alternative source of information for distinguishing family developments. This process enables us to assess the accuracy of the target population variable in the HUD database on populations actually served and to assess the performance of our methodology for identifying family developments.

Methodology

Our basic approach is to exploit differences between the distribution of unit sizes (where size is the number of bedrooms in a unit) in family versus nonfamily developments among those developments for which we have very good information on the population served, namely those developments with good data on a targeted population. We use those observed differences to develop an algorithm for identifying family developments within the remaining developments, those for which the target population is not known. For our approach to work, observable differences need to exist in the unit sizes found in family and nonfamily developments, which is testable in the national data. These differences also need to hold for family developments that have *not* been identified as such in the HUD data, which cannot be tested with publicly available data. Using our tenant-level data for six states, however, we can assess whether our algorithm does a good job at capturing family developments among developments for which the target population is incomplete in the HUD data, at least in those six states. This method provides an "out-of-sample" assessment of the algorithm and also a method for assessing the validity of the HUD variables on the target population.

Those developments for which we have the best evidence that they are or are not serving families are those identified in the HUD data either as explicitly targeting families (family) or those

[6] LIHTC target population categories include family, elderly, homeless, disabled, and other. These categories need not be mutually exclusive, although some states indicate only one.

[7] Officials from several state housing finance agencies (the allocating agencies for the LIHTC Program) reported that developments serving families are coded as "nontargeted" in their state.

[8] We also supplement the HUD LIHTC data with project data from one state.

[9] These data are part of a larger LIHTC project, using data from more than 30 states. Here we focus on the 6 states that provide individual-level (rather than household-level) data, including age, potentially permitting us to identify the presence of children.

identified as targeting the elderly (nonfamily).[10] Nationally, 35 percent of developments can be identified as targeting either families or the elderly (12,848).[11] Within this group, 84 percent (10,772) have complete information on units by number of bedrooms. These developments include 691,331 units, and this sample is used to determine the algorithm.

The table in exhibit 1 provides information on the units by number of bedrooms within these two types of developments. As expected, a much larger share of units in family developments are multibedroom. More than 75 percent of units in family developments have at least two bedrooms, while the opposite is true for nonfamily (elderly) developments—where more than 75 percent of units have fewer than two bedrooms. Indeed, other researchers have used this difference to classify large units (two bedrooms or more) as family units. Although 77 percent of the units in family developments were correctly captured, this proxy misses 23 percent of the units actually in family developments (type I error).[12] Of units ultimately labeled as being in family developments, 13 percent would in fact be in elderly developments (type II error). This seems a fairly good proxy, particularly given its ease of application. The question at this point is whether we can improve on this proxy by using information on the full distribution of units by number of bedrooms or whether we can provide an alternative proxy for researchers interested in focusing on family developments rather than units, one that performs at least as well in accuracy.

Exhibit 1

Unit-Size Distribution of Developments by Target Population

Unit Size	Total Units		Share of Units of Each Size (%)	
	Nonfamily/Elderly Developments	Family Developments	Nonfamily/Elderly Developments	Family Developments
0 bedrooms	10,972	14,050	5	3
1 bedroom	152,415	95,376	71	20
2 bedrooms	47,643	213,522	22	45
3 bedrooms	4,520	129,884	2	27
4 bedrooms	381	22,568	0	5
Total	215,931	475,400	100	100

Algorithm: In Brief

Exhibit 1 reveals some noticeable differences in the unit-size distributions: very large units (three or four bedrooms) are nearly exclusively in family developments, while large concentrations of one-bedroom units and small shares of two-bedroom units are primarily in nonfamily, elderly developments. These differences are the type we exploit to define developments as either family or nonfamily. In addition, we develop our algorithm through a series of sequential classification steps. After each step, we reexamine the unit-size distributions of the remaining family and nonfamily developments,

[10] Families may also be served in the remainder of developments, which we return to in our out-of-sample test of algorithm.

[11] The HUD database identifies 12,375 developments as targeting families or the elderly. We supplemented with project data from one state poorly covered in the HUD data to reach 12,848 developments.

[12] We assume this method is meant as a proxy for units in family developments; later we discuss this method as a proxy for units housing children.

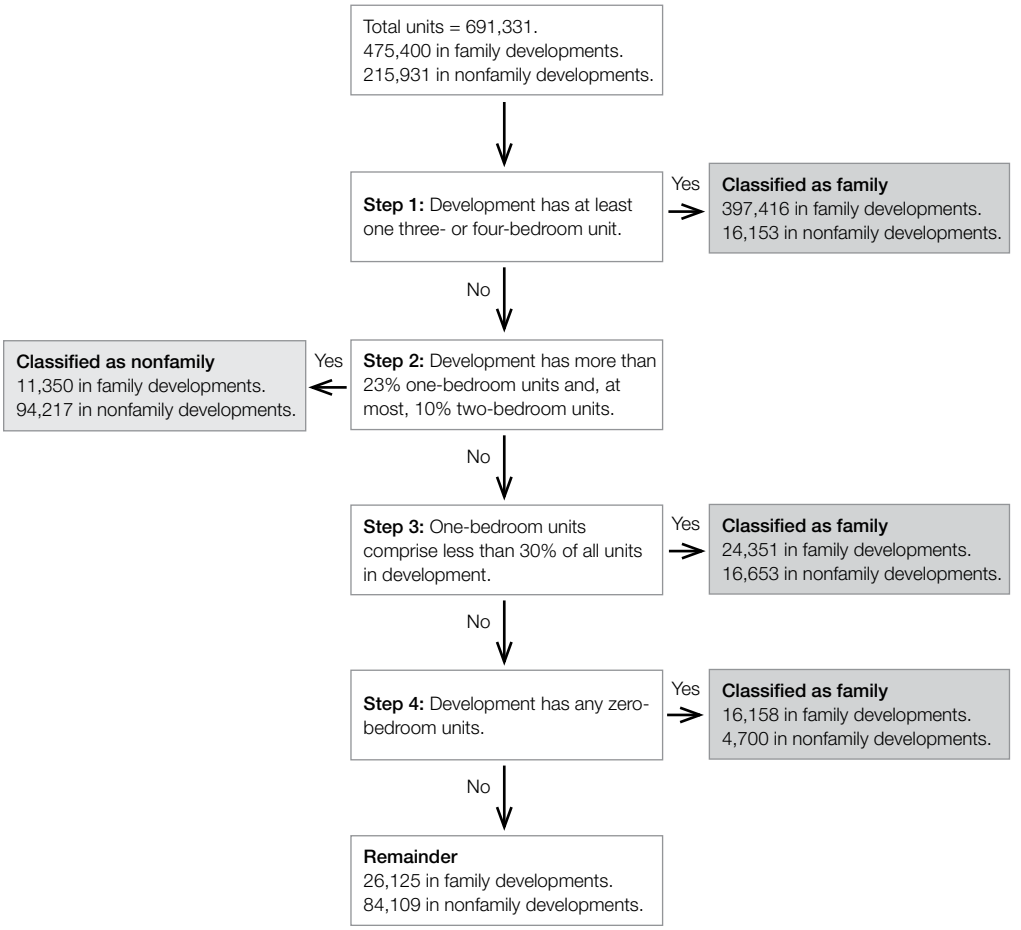
those that have not yet been categorized, to tailor additional categorization criteria so as to capture the greatest share of units in family developments while minimizing the number of units in non-family developments misclassified. We have gone through numerous iterations and assessments of type 1 and type 2 errors. Our preferred algorithm has four steps.

Algorithm: The Details

Exhibit 2 displays a flowchart that illustrates how developments are classified as either family or nonfamily during each iteration within the algorithm. The algorithm uses the 691,331 units in 10,772 developments placed in service through 2010 that target either families or the elderly and for which complete information is available on the size of units in the development. Examining the distribution of units within developments by target population (exhibit 1) revealed that very large units are primarily located in family developments. This observation produced step 1 of the algorithm.

Exhibit 2

How Developments Are Classified As Either Family or Nonfamily During Each Iteration Within the Algorithm



Step 1. Developments that have at least one three- or four-bedroom unit are classified as family developments.

This first step correctly identifies 84 percent of units in family developments while misclassifying 8 percent of units in elderly developments. (See the table in appendix exhibit A-1 for details.) We then examined the distribution of units within the remainder group and identified that large shares of small units combined with low shares of large units primarily occur in nonfamily, elderly developments. Specifically,

Step 2. Developments with more than 23 percent one-bedroom units and, at most, 10 percent two-bedroom units are assigned to the nonfamily category.

This step correctly identifies 44 percent of all elderly units and misclassifies 2 percent of all family units. Although our focus is on identifying family developments, the removal of developments that are recognizable as nonfamily (elderly) decreases type 2 errors in later steps. Again examining the unit-size distribution for remainder developments, we find that family developments don't contain large shares of one-bedroom units. Specifically,

Step 3. Developments where one-bedroom units comprise less than 30 percent of all units are assigned to the family development category.

This step correctly identifies 5 percent of the family units and misclassifies 8 percent of the elderly units. In examining the unit-size distribution for the remaining 131,000 units separately by target category, we discovered something counterintuitive; among the remaining developments, *family* projects are more likely to contain a studio apartment than are elderly developments (exhibit 3). Indeed, while 94 percent of the remaining elderly developments contain no studios, more than 38 percent of the remaining family units contain at least one studio.[13]

This surprising result is driven by step 2, which removes well over 40 percent of elderly developments from the sample based on large concentrations of small units. Those senior developments

Exhibit 3

Share of Units With Zero Bedrooms

Percentiles	Nonfamily/Elderly Developments	Family Developments
1	0.00	0.00
60	0.00	0.00
70	0.00	0.10
75	0.00	0.14
90	0.00	0.32
95	0.01	0.38
99	0.26	0.42
N	88,908	42,283

[13] Of course, we cannot rule out that some of these family developments with large shares of studios are in fact misclassified in the HUD data, but our assessment using age of heads of household suggests that the number of such misclassifications is quite small.

that have any studios also have large concentrations of one-bedroom units, so they have already been classified as nonfamily. This scenario highlights the benefit of reassessing the size distribution of developments between each step.

Step 4. Developments with any zero-bedroom units are assigned to the family development category.

This step correctly identifies another 3 percent of the family units and misclassifies 2 percent of elderly units.

The table in exhibit 4 provides a summary of how units are classified in the HUD data versus the algorithm. (A more detailed table is provided in appendix exhibit A-1.)

Exhibit 4

Summary of Family Classification Outcomes

	Family Developments	Nonfamily/Elderly Developments	Total
Units according to HUD data	475,400	215,931	691,331
Units classified as family through algorithm	437,925	**37,605**	475,530
Units not classified as family through algorithm[a]	**37,475**	178,326	215,801

HUD = U.S. Department of Housing and Urban Development.
[a] At the completion of the algorithm, not all units will be classified. Units not classified as family include those classified as nonfamily (step 2) or not classified as family at any point in the algorithm.
Note: Bold indicates incorrect family classification.

Discussion

This algorithm correctly classified 92 percent (437,925) of units in family developments. Alternatively, 8 percent (37,475) of the family units are not classified correctly (type I error) and only 8 percent (37,605) of the units classified as being in family developments are actually in nonfamily developments. In terms of developments, 90 percent of the family developments are correctly classified and 9.6 percent of developments classified as family are incorrectly classified.

This algorithm appears to perform quite well within a sample of developments that are identified as either family or nonfamily (elderly). Although the algorithm is promising, we would also like some sense of how well it would work outside this sample; that is, on the group of developments without clear information on the population served—the population to which it would actually be applied.

To help assess this classification scheme, we rely on tenant-level data from six states, which provide information on the age of members of the household, helping to identify children and the elderly. While data for the head of household are nearly always complete, coverage for additional members unfortunately is not, which limits our ability to capture the presence of children to two states. Given that the primary form of nonfamily developments is for the elderly, however, we have an alternative option of identifying developments in which disproportionate share of households are headed by a senior.

Presence of the Elderly

Our first step is to assess whether data on the presence of the elderly can adequately distinguish family from nonfamily developments. For family and nonfamily developments separately, we calculate the share of households in a development in which the head of the household is 55 years old or older.[14] Exhibit 5 presents the distribution.

The two distributions in exhibit 5 are quite different. Although more than 95 percent of elderly developments have at least 50 percent of their households headed by a senior citizen, only 5 percent of family developments do. This difference suggests two things. First, the LIHTC variables on whether developments target the elderly and families appear quite good. Second, for the six states for which we have data on the age of the head of the household, using that data should provide a good alternative means of determining the populations housed by developments. We use these data specifically to informally assess the accuracy of the bedroom algorithm to categorize developments that are not identified in HUD data as targeting families or the elderly.

Exhibit 5

Distribution of the Share of Units With Household Heads Age 55 or Older (six-state sample)

Percentiles	Family Developments	Nonfamily/Elderly Developments
1	0.00	0.20
5	0.04	0.63
10	0.06	0.79
25	0.11	0.92
50	0.18	0.99
75	0.28	1.00
90	0.43	1.00
95	0.53	1.00
99	0.98	1.00
N	67,626	37,320

Out-of-Sample Test

In the six states, we applied our classification algorithm to those developments not previously identified as either family or the elderly. This group contains three types of developments: those that target some other group (that is, homeless), those that are labeled as "not targeting," and those for which simply no information exists on targeting. We then examined the distribution of the share of households headed by the elderly in these developments, now classified as either family or nonfamily (the elderly), presented in exhibit 6.

The distribution for family developments in exhibit 6 looks very similar to the distribution in exhibit 5. Slightly less than 95 percent (94 percent, number not in exhibit) of developments now classified as family have less than 50 percent of units with household heads who are 55 years old

[14] We also looked at the presence of any elderly in the household and defined elderly as 62 years old or older. Results are similar, but focusing on heads of household and 55 years old or older provides the largest differences between family and elderly developments.

Exhibit 6

Share of Units With Household Heads Age 55 or Older, by Classification

Percentiles	Classified Family	Classified Nonfamily
1	0.00	0.00
5	0.03	0.09
10	0.05	0.14
25	0.08	0.48
50	0.12	0.71
75	0.19	0.95
90	0.32	0.99
95	0.50	1.00
99	1.00	1.00
N	20,531	3,209

or older. The nonfamily distribution varies more from that in exhibit 5, but the absolute number of nonfamily developments is quite small, which indicates that to the extent that this variation reflects misclassification, this error should be small in magnitude.

To more systematically assess the performance of the algorithm, we use the "50 percent headed by the elderly" as a firm cutoff, which indicates that we assume developments below that threshold are actually family developments, and those above it are actually elderly developments. Given this assumption, we can then assess how well the algorithm performs. Exhibit 7 provides a summary of the results, first for all units in the sample (column 1), then broken out for subgroups based on how the developments are categorized in the HUD data.

Overall, the algorithm is estimated to have correctly classified 96 percent of units in family developments (4 percent type I error), with a 6-percent error rate among units so classified (type II error). As an additional check, in the two states for which we have complete data on children in the household, we find that 98 percent of children in LIHTC housing are located in developments identified as family by this method.

Exhibit 7

Applying the Algorithm Out of Sample (six states) by Development Classification

	All	No Information	Not Targeted	Other Target
Total units	23,740	10,824	6,599	6,317
Percent of units classified as in family developments when using:				
Share of household heads who are elderly as the cutoff	85	77	88	94
Bedroom algorithm	86	77	95	95
Algorithm performance (assuming household heads who are elderly is correct):				
Percent type I errors	4	7	1	4
Percent type II errors	6	7	8	5

Exhibit 7 also reveals that most developments not classified as targeting family or the elderly in the HUD data are indeed family developments, according to the distribution of number of bedrooms, the age of the head of the household, and the presence of children (where we can assess this). To analyze only those developments with complete data on targeted populations disproportionately misses family developments.

Caveats and Conclusion

For researchers interested in identifying family developments, employ a classification scheme based on information on the full distribution of bedrooms in such developments (in contrast to nonfamily developments), which permits a much more comprehensive assessment of the LIHTC stock. This method, of course, will include errors. Our best assessment of the error rate suggests that our algorithm does a very good job of correctly classifying actual units in family developments when applied to developments for which we have incomplete information on their target population. In the six states, we estimate this approach captures approximately 96 percent of the units in family developments (90 percent of family developments), with single-digit type II error rates, very similar to the error rates within the sample of developments clearly identified as family or nonfamily in the HUD data nationally. Whether those estimated error rates are acceptable will depend on the goal of the work, but it does provide a much more complete coverage of LIHTC developments and one at a national scale.

For researchers focused on smaller regions or states, we suggest that a similar approach be taken, but that it be tailored to the geography. We found some variation across the states in the distribution of number of bedrooms for family and elderly developments in the HUD data. Researchers can exploit that variation by devising their own algorithm—one that performs best for the particular state or region.

Finally, for researchers specifically interested in identifying units most likely to house children (rather than developments of families more broadly), we did some additional assessment of where *current* children live, by unit size, in the two states for which we have complete data on children. Approximately 98 percent of children live in units that have two bedrooms or more. This unit-size proxy does a remarkably good job at identifying units likely to house children (very low type I error). The proxy does not avoid units that do not house children (type II error), however. In those two states, approximately 40 percent of large units do not currently contain children. Of course, those units may house children at another point in time. Large units that are in elderly developments will not house children at any point, however. Exhibit 1 suggests that the large-unit proxy has a type II error rate at least in the double digits. Researchers interested in a unit-based proxy that focuses on children rather than families would benefit from combining the two approaches; that is, they would use the development algorithm to remove the elderly developments and thereby all large units in the elderly developments, which likely are the greatest source of error for the unit-based approach.

Appendix

Exhibit A-1

Algorithm Classifications by HUD Target Population

		Units				Developments			
		Family		Nonfamily (Elderly)		Family		Nonfamily (Elderly)	
	Classification	Total	Share of Total (%)	Total	Share of Total (%)	Total	Share of Total (%)	Total	Share of Total (%)
To start		475,400	100	215,931	100	7,306	100	3,466	100
Step 1	Family	397,416	84	16,153	8	5,684	78	278	8
Step 2	Nonfamily	11,350	2	94,217	44	210	3	1,555	45
Step 3	Family	24,351	5	16,653	8	774	11	356	10
Step 4	Family	16,158	3	4,799	2	110	2	62	2
	Remainder (not classified)	26,125	6	84,109	39	528	7	1,215	35
Final results	Total classified as family	437,925	92	37,605	17	6,568	90	696	20

HUD = U.S. Department of Housing and Urban Development.

Acknowledgments

The authors thank the National Council of State Housing Agencies and their numerous members who voluntarily contributed data to the Furman Center for Real Estate and Urban Policy. They also thank Ingrid Gould Ellen and Keren Mertens Horn for their helpful feedback on earlier drafts. All remaining errors are the authors' own.

Authors

Rachel M.B. Atkins is a doctoral student at the New School.

Katherine M. O'Regan is a professor of public policy and planning at the Wagner Graduate School, New York University.

References

Ellen, Ingrid, and Keren Horn. 2012. *Do Federally Assisted Households Have Access to High Performing Public Schools?* New York: Furman Center for Real Estate and Urban Policy. http://furmancenter.org/files/publications/PRRACHousingLocationSchools.pdf.

Kawitzky, Simon, Fred Freiberg, Diane L. Houk, and Salimah Hankins. 2013. *Choice Constrained, Segregation Maintained: Using Federal Tax Credits to Provide Affordable Housing*. New York: Fair Housing Justice Center. http://www.fairhousingjustice.org/wp-content/uploads/2013/08/FHJC-LIHTCREPORT-Aug13-Fullv1-7-WEB.pdf.

Khadduri, Jill, Larry Buron, and Carissa Claminco. 2006. *Are States Using the Low Income Housing Tax Credit to Enable Families with Children to Live in Low Poverty and Racially Integrated Neighborhoods?* Washington, DC: Poverty and Race Research Action Council. http://www.prrac.org/pdf/LIHTC_report_2006.pdf.

Khadduri, Jill, Larry Buron, and Ken Lam. 2004. "LIHTC and Mixed Income Housing: Enabling Families with Children to Live in Low Poverty Neighborhoods?" Paper presented at The Association of Public Policy and Management 26th Annual Research Conference, October 30. http://www.abtassoc.net/reports/Khadduri_%5B7%5D_LIHTC_Mixed_Income_APPAM.pdf.

Newman, Sandra, and Ann Schnare. 1997. "…and a Suitable Living Environment: The Failure of Housing Programs To Deliver on Neighborhood Quality," *Housing Policy Debate* 8 (4): 703–741.

Pfeiffer, Deidre. 2009. "The Opportunity Illusion: Subsidized Housing and Failing Schools in California." Los Angeles: The Civil Rights Project. http://civilrightsproject.ucla.edu/research/metro-and-regional-inequalities/housing/the-opportunity-illusion-subsidized-housing-and-failing-schools-in-california/pfeiffer-opportunity-illusion-2009.pdf.