Appendix 7.a

TECHNICAL AND METHODOLOGICAL ISSUES

Calculation of Segregation Index

For a given district made up of *j* attendance zones, in time *t*, Clotfelter utilizes the following index measure of segregation: $S_t = \frac{n_t - \left(\sum_{j=1}^{n} W_{jt} n_{jt} / \sum_{j=1}^{n} W_{jt}\right)}{n_t}$.[1] Here, for time *t*, n_t represents the proportion of residents who are non-white in the district, as a whole; W_{jt} represents the number of white residents in attendance zone j; and n_{jt} represents the proportion of residents who are non-white in attendance zone j. The quotient within the parentheses represents the overall exposure rate between whites and non-whites in the district. Since this value is sensitive to the overall proportion of non-whites, Clotfelter standardizes the exposure rate by the overall proportion of residents in the district who are non-white to generate a segregation index.

Modeling Families' Residential Preferences

Our analytic goal is to understand whether and the extent to which families' revealed preferences for racially homogenous school attendance zones changed in the aftermath of the unitary status declaration. To assess this change in revealed preferences, we examine, among families who move, year to year changes in the likelihood of families selecting into a school attendance zone that is more similar to their child's own race than the zone that they depart. To do so, we require an analytic approach that allows us to describe individual household choice as a function of characteristics that are specific to the combination of the household and of each possible option from which a household can choose. We therefore utilize McFadden's conditional logit model, which allows us to examine the factors that govern a family's decision not only of *whether* to move but also of *where* to move.[2]

We first outline in formal terms the theoretical framework for why the conditional logit model properly describes the residential choices families will make. Then, we describe how we format our dataset to permit estimating the conditional choice model. Finally, we describe the model itself.[3]

Assume that a given family i, has j school attendance zones from which to choose and that each school zone can be described by a vector of characteristics Y_j. These characteristics might include average property value, school quality, local amenities, proximity to public transportation, and demographic (e.g., racial) make-up of the zone residents. Let X_i represent family characteristics such as race and prior school achievement of the children in the household. The value of the j^{th} attendance zone to family i is $U(Y_j,X_i)$. U denotes utility, and $U(Y_j,X_i)$ indicates that the utility that family i would gain from residing in attendance zone j is a function of the characteristics both of attendance zone j and of family i. Following these definitions,

$$U(Y_j,X_i) = E(Y_j,X_i) + \varepsilon_{ij},$$

where, $E(Y_j,X_i)$ represents the mean utility of Y_j for individuals with a vector of characteristics X_i, and ε_{ij} represents the random variation among families that depends on unobservable preferences.

We assume that the non-random portion of a family's utility for a particular zone is a function of that school zone's characteristics and the interaction between school zone and household-level characteristics. These interactions represent household-zone specific measures. In contrast, household characteristics on their own are not included in considering utility for particular school zones, as a family's characteristics, in a vacuum, should not influence choice of residence. Rather, it is only how a family's characteristics match a neighborhood of potential residence that should have an effect on whether or not a family selects a given attendance zone. We highlight below why this point is important from an analytic perspective.

We assume that for each household, school zone selection will be utility maximizing, subject to the household's budget constraint. That is, family i selects Y_k if and only if:

$U(Y_k, X_i) \geq U(Y_j, X_i)$ for all $k \neq j$, subject to the household budget constraint.

Therefore, our model considers each family's choice among the j potential school attendance zones. To fit our model, we organize the data as pair-wise combinations of each family i with each school attendance zone j, for a total of i×j observations. While the number of schools (and associated school attendance zones) varied somewhat from year-to-year, organizing the data in this way in each year yields between 67 and 78 observations for each family with an elementary-aged child.[4]

Having organized the data in this way, the model that we specify is made up of j equations for each family i with each equation describing one of the elements (i.e., zones) in the choice set. In fitting this model, we estimate the probability of each family i choosing to live in school zone j, relative to all other alternatives, in year t.[5] The outcome, $ZONE_{ijt}$, is equal to one for the school zone actually chosen by family i in year t and zero for all other zones. This allows us to model explicitly the tradeoffs between the school zone selected and the unselected alternatives.[6] The primary predictor variable interacts $LESS\_WHITE_{ijt}$, which is equal to one for each zone in which a higher proportion of the CMS students residing within it are non-white compared to the student's current zone of residence and zero otherwise. *YEAR*, represents a linear time trend re-centered on 0 in the year 2002, the year of the unitary status declaration. Finally, $POST_t$ is equal to one in 2003 and subsequent years and indicates the years after the declaration of unitary status.[7]

Our research question asks whether the unitary status declaration caused families to make segregative moves; however, the change in assignment policy permitted families to control the makeup of their children's school according to observable school-level characteristics other than racial make-up. In order to differentiate moves that reflect a preference for more racially segregated neighborhoods and ones that reflect a preference for higher achieving schools, we also introduce a critical control variable, HI_ACH_{ijt}. HI_ACH_{ijt} is equal to one if the school associated with a particular zone has average standardized math and reading achievement scores that are higher than the student's initial zone of residence.[8] To capture the fact that one choice available to families is to not move, we capitalize on Mare and Bruch's (2003) strategy and include the control variable $STAY_{ijt}$ which is equal to one for the school assignment zone in which family i initially lived, and zero otherwise. The inclusion of *STAY* also permits the conditional choice model to capture the non-linear jump from a family deciding whether to move

as opposed to deciding where to move. In certain specifications, we also include the student-level variable, $ACHIEVE_{ijt}$, a student's performance on the North Carolina End-of-Grade mathematics assessment, to detect whether families with children with higher academic performance might be more motivated to move in search of more homogenous schools and neighborhoods. [9]

The zone level variables *STAY*, *LESS_WHITE*, and *HI_ACH* in our model correspond to the characteristics of zone Y_j in Equations (2) and (3). Where *MOVE* is equal to 1, these characteristics are defined as Y_k. The interaction between these zone-level features and individual characteristics such as race and *ACHIEVE* equate to the household-zone characteristics of X_i. Thus, in our simple specification, the utility of a residence for a given family is a product of its zone-level racial and school characteristics, the zone characteristics interacted with the child's racial and achievement profile, and family- and child-specific unobservables.

For the sake of clarity in writing out the model below, we represent the control variables *STAY*, *HI_ACH*, and *ACHIEVE* and their interactions with each other and with *LESS_WHITE* as $\mathbf{C}_{ijt}$. $(\mathbf{C}\times\mathbf{T})_{ijt}$ represents a vector of the interaction between vector $\mathbf{C}_{ijt}$ and time variables, *YEAR*, *POST* and *POST*×*YEAR*. We fit the following conditional logit choice model:

$$P\left(ZONE_{ijt+1}\right) = \frac{exp^{Z_{ijt}\beta}}{exp^{Z_{i1t}\beta} + exp^{Z_{i2t}\beta} + \cdots + exp^{Z_{iJt}\beta}}, \text{ where}$$

$$Z_{ijt}\beta = \beta_1 LESS_{WHITE\,ijt} + \beta_2 LESS_{WHITE} \times YEAR_{ijt} + \beta_3 LESS_{WHITE} \times POST_{ijt} +$$

$$\beta_4 LESS\_WHITE \times POST \times YEAR_{ijt} + \mathbf{C}_{ijt}\gamma + (\mathbf{C}\times\mathbf{T})_{ijt}\delta + \varepsilon_{ijt}$$

The parameters of interest are β_3 and β_4, and their estimates will be negative and statistically significant for whites if the policy shift caused an increase in segregative zone choices among families who moved.

We underscore that both *LESS_WHITE* and *HI_ACH* are binary measures. While we recognize that this modeling choice results in some loss of granularity, we choose to use binary variables in our primary specifications because our central interest is in whether families are more likely after unitary status to make segregative moves and not in the more nuanced question related to the functional form of the relationship between zone selection and the size of differentials in racial makeup. Given this goal, the binary variables allow for clear and more easily interpretable results with respect to the key question of whether the policy change increased the likelihood of segregative residential movement.

As noted above, for each consecutive pair of years (t, t+1), this analytic procedure involves estimating a set of j equations for each family i, as for each family we are interested in the probability of selecting from among a set of j options. As a consequence of estimating within each family, student- and family-level characteristics do not enter the equations as main effects. Rather, they enter as interactions with the characteristics of specific choices. Therefore, in order to incorporate student-level racial characteristics, we simply estimate models separately for whites and non-whites. Given that our data includes a near-census of all students served by CMS, subsetting the data in this way does not unduly threaten the precision of our estimates.

Ideally, we would define time-specific alternatives and person-period cases in the conditional logit model and then cluster standard errors at the student level to reflect the fact that some households are observed in the data for several years. Without these additions, the model is already computationally intensive. Unfortunately, when we attempt to fit this augmented specification, the model fails to converge. Therefore, we treat each period-specific observation of an individual as independent from any other-period observation of that person. Specifying the model in this way treats each individual's probability of moving in a given year as independent from choices made in any prior year. This simplifying assumption means that our residuals will necessarily be correlated, leading to an underestimate of our parameter estimate standard errors. To address this concern, we conducted sensitivity analyses to explore the extent to which our standard errors were underestimated. In order to do so, within each family's set of observations (each corresponding to a possible move from year t to year t+1), we randomly sampled a single observation and refit our model with this reduced sample. While it would be feasible, technically, to use this approach to generate standard errors empirically (similar to bootstrapping), this was impossible given the computational demands of fitting even a single iteration of the conditional logit model with a dataset of this size. Therefore, we repeated this procedure ten times in order to gauge the extent to which we should inflate our standard errors. This sensitivity check also yielded point estimates that were essentially unchanged our primary results, providing assurance that the repeated observation of children overtime did not lead to bias in our point estimates.

---

[1] Charles H. Clotfelter, *After Brown: The Rise and Retreat of School Desegregation* (Princeton, NJ: Princeton University Press, 2004).

[2] Daniel McFadden, "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in econometrics*, ed. Paul Zarembka (New York: Academic Press, 1973), 105-142; Daniel McFadden, "Quantitative Methods for Analyzing Travel Behavior of Individuals: Some Recent Developments," in *Behavioural Travel Modeling*, ed. David A. Hensher and Peter R. Stopher (London: Croom-Helm, 1979).

[3] The following section relies extensively on: Long's excellent explanation of the conditional choice model in her analysis of college-going patterns in the last quarter of the 20th century (Bridget Terry Long, "How Have College Decisions Changed Over Time? An Application of the Conditional Logistic Choice Model," *Journal of Econometrics* 121, no. 1-2 (2004): 271-296); Mare and Bruch's analysis of residential mobility and segregation in Los Angeles (Robert D. Mare and Elizabeth E. Bruch, "Spatial Inequality, Neighborhood Mobility, and Residential Segregation" (working paper, California Center for Population Research On-Line Working Paper Series 002-03, University of California Los Angeles, 2003); and Bruch and Mare's examination of the extent to which individuals respond to the racial makeup of their neighborhoods (Elizabeth E. Bruch and Robert D. Mare, "Neighborhood Choice and Neighborhood Change," *Journal of Sociology* 112, no. 3: 667-709.

[4][4] A simplifying assumption that we make is to ignore the presence of siblings in the data set. This could potentially lead to bias in our results as a consequence of residuals that are correlated

across children within families. The data do not include information on sibling pairs. As a sensitivity check, we identify presumptive siblings by linking those students with the same last name and home address across several years of data. Among sets of presumptive siblings, we then retain only one child and rerun our analyses. Both point estimates and standard errors associated with these sensitivity analyses are largely unchanged, and substantive conclusions remain the same. Given the robustness of our results to this sensitivity check, combined with a concern that the quality and accuracy of the matches (given the possibility of multiple last names among siblings), we nevertheless prefer the full sample results. Results from this sensitivity check are available upon request. Additionally, we minimize overestimation from siblings within each set of models by estimating results separately for elementary, middle and high school zones. We address correlated residuals in more detail below.

[5] Year represents the spring semester residence, so that when we report results for 2002 they reflect the probability of moving between the spring semester of the 2001-2002 school year and the spring of the 2002-2003 year.

[6] Stata includes a routine to estimate McFadden's conditional logistic choice model: ASCLOGIT. In our application, the case() option represents individuals while the alternatives() option represents the school attendance zones.

[7] Of course, a linear specification of time imposes a functional form constraint on this variable. While not presented here, we first fit a completely general specification of time, with dummy variables for each year. The results from this model indicate that a linear specification of time is reasonable.

[8] We assess school-level achievement using average student performance on the North Carolina End-of-Grade assessments in reading and mathematics in grades 3 – 8 and 10, standardized to mean 0, standard deviation of 1.

[9] Correlations between scores on End-of-Grade reading and mathematics exams are well above .90. As expected, therefore, results based on the inclusion of this measure of student achievement are not sensitive to the choice of the mathematics, as opposed to reading, score.