DOCUMENT RESUME

ED 171 825 UD 019 420

AUTHOR	Becker, Henry Jay
TITLE	The Measurement of Segregation: The Dissimilarity Index and Coleman's Segregation Index Compared. Report No. 265.
INSTITUTION	Johns Hopkins Univ., Baltimore, Md. Center for Social Organization of Schools.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Nov 78
GRANT	NIE-G-78-0210
NOTE	40p.
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Comparative Analysis; Data Analysis; *Integration Studies; *Measurement Instruments; *Racial Segregation; *Statistical Data; Statistical Studies
IDENTIFIERS	*Coleman Segregation Index; *Index of Dissimilarity

ABSTRACT

Most measurements of racial and ethnic segregation, particularly comparative analyses across cities, have relied on the Index of Dissimilarity ("D"). Recently, however, it has been demonstrated that cities with different racial compositions also have different expected values of "D" under a random distribution of whites and blacks. Here "D" is compared with another, increasingly used, segregation index which we call "S". Applied to measurement of school segregation, "S" equals the proportional underrepresentation of black students in the school attended by the average white student in the district. A principal value of "S" is that its expected value is independent of racial composition and rapidly approaches zero, even for modest-sized units of analysis. Also, "S" can be "decomposed" into segregation within subsets of entities and segregation between subsets. The behavior of "S" and "D" are explored with data on segregation in higher education and segregation of workers across different places of employment. (Author)

THE MEASUREMENT OF SEGREGATION: THE DISSIMILARITY INDEX AND COLEMAN'S SEGREGATION INDEX COMPARED

Grant No. NIE-G-78-0210

Henry Jay Becker

Report No. 265

November 1978

Published by the Center for Social Organization of Schools, supported in part as a research and development center by funds from the United States National Institute of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the Institute should be inferred.

The Johns Hopkins University

Baltimore, Maryland

Introductory Statement

The Center for Social Organization of Schools has two primary objectives: to develop a scientific knowledge of how schools affect their students, and to use this knowledge to develop better school practices and organization.

The Center works through four programs to achieve its objectives. The Policy Studies in School Desegregation program applies the basic theories of social organization of schools to study the internal conditions of desegregated schools, the feasibility of alternative desegregation policies, and the interrelation of school desegregation with other equity issues such as housing and job desegregation. The School Organization program is currently concerned with authority-control structures, task structures, reward systems, and peer group processes in schools. It has produced a large-scale study of the effects of open schools, has developed Student Team Learning Instructional processes for teaching various subjects in elementary and secondary schools, and has produced a computerized system for school wide attendance monitoring. The School Process and Career Development program is studying transitions from high school to post secondary institutions and the role of schooling in the development of career plans and the actualization of labor market outcomes. The Studies in Delinquency and School Environments program is examining the interaction of school environments, school experiences, and individual characteristics in relation to in-school and later-life delinquency.

This report, prepared by the Policy Studies in School Desegregation program, furthers the scientific measurement of racial and ethnic desegregation by comparing the utility of two measures--the Index of Dissimilarity and the Segregation Index.

Abstract

Most measurements of racial and ethnic segregation--particularly comparative analyses across cities--have relied on the Index of Dissimilarity ("D"). Recently, however, it has been demonstrated that cities with different racial compositions also have different expected values of D under a random distribution of whites and blacks.

Here we compare D with another, increasingly used, segregation index which we call "S". Applied to measurement of school segregation, S equals the proportional underrepresentation of black students in the school attended by the average white student in the district. A principal value of S is that its expected value is independent of racial composition and rapidly approaches zero, even for modest-sized units of analysis. Also, S can be "decomposed" into segregation within subsets of entities and segregation between subsets.

The behavior of S and D are explored with data on segregation in higher education and segregation of workers across different places of employment.

Acknowledgments

This paper was written with the assistance of James M. McPartland, Gary A. Chase, Denise C. Daiger and Gail E. Thomas.

Meanings of "Segregation"

The term "segregation" has a variety of specific meanings both in everyday discourse and in social scientific and policy research. Most often, it is used to describe a single establishment or environment (that is a school, a workplace, a neighborhood).

For example, a place may be considered segregated if members of a socially salient category constitute either an overwhelmingly large proportion of the members or inhabitants (e.g. 95% black) or an exceedingly small proportion (e.g., 5% black). The significant feature of such a place is that both the experience of living in homogeneous situations and the experience of life as part of a tiny "minority" have important consequences for individual attitudes, social actions or structural events.

Another use of the term "segregation" draws attention to a specific place that deviates markedly from the average group membership composition among a set of places in a certain universe. For example, a place may be considered segregated if its racial composition is 50% black and 50% white because the places in this particular universe average only 20% black and 80% white. Here the significant feature is the inequality among places in the experience of group heterogeneity. The "segregated" places are those which most deviate from the average experience in the universe of places.

In contrast to both of these uses of the term "segregation," the word is also applied to a collection or aggregate of collective units

such as a school district, a city divided into census tracts, or all the places of businesses within a particular industry. When so used, the emphasis is usually on the degree of variation among the units within the aggregate. For example, the schools in city A may be considered more segregated than city B's if there is more variation in the "percent black" among the schools in city A than among the schools in city B. City A may have a more ethnically heterogeneous total enrollment than city B, but the distribution of pupils from the various ethnic groups across the schools in city A is less uniform.

Generally, the use of the term "segregation" in social scientific research has been closer to this last definition than it has been to the others. For more than a decade, a single measure of segregation, the "dissimilarity index," has dominated the research literature on segregation.

The Index of Dissimilarity

The dissimilarity index, D, is based on the unevenness of the univariate distributions of two groups (e.g., blacks and non-blacks) among units such as schools or places of business. With the dissimilarity index, the measurement of unevenness is represented by a Lorenz curve.

To understand D and the Lorenz curve, consider a set of mutually exclusive collective units like schools or places of business. First, array each of these units in ascending order according to the proportions of their members (students, employees) who are in a particular group (e.g., their percent black). A Lorenz curve, in this case, is

given by graphing the cumulative proportion of all non-blacks in the aggregate who belong to those units considered up to that point against the cumulative proportion of all blacks in those same units. (Figure 1 illustrates this curve.) In a completely integrated aggregate, if a unit contained 10% of all the people in category X, it would also contain 10% of all of the category $\overline{X}$'s (non-X's). Under segregation, units that are "overrepresented" among one group's population are necessarily underrepresented among the residual group.

The value of the dissimilarity index is given by the vertical distance between the Lorenz curve and the point along the 45^{o} diagonal (representing a completely integrated aggregate) at the point where the curve is tangent to a line parallel to the 45^{o} line (see Figure 1). This vertical distance represents the accumulated underrepresentation of one group's population up to the point where the next unit added is the one which has a proportionate allocation of members to the two groups considered. Thus, the dissimilarity index is the minimum proportion of members of a given group, presently located in units "underrepresented" in that group, who would have to move or be transferred to units where such persons are overrepresented in order for each unit to contain the identical fraction of each groups' population.

Where one is dealing with two exhaustive groups (such as blacks and non-blacks), D is given by:

$$D = \frac{\sum_i n_i \left| p_i - P \right|}{2 N \cdot P (1-P)},$$

where n_i = total persons, i'th unit;
p_i = percent black, i'th unit;
P = overall percent black;
N = total persons

Because D is based on a single point in the bivariate cumulative distribution (the point of proportionate representation among the ordered units), it does not distinguish between various degrees of segregation among the "overrepresented" or among the "underrepresented" units. In Figure 1, the dashed line represents a different universe where no one is found in an exclusively black or non-black unit, but which has the same D score as the universe shown by the solid line.

The concept of dissimilarity and its measurement are not limited to situations of exhaustive dichotomies. The same notion of under- and overrepresentation can be applied to any two mutually exclusive groups (e.g., Hispanics and blacks). Under these circumstances D has the following algebraic representation:

$$D = 1/2 \sum_i \left| \frac{H_i}{H} - \frac{B_i}{B} \right|,$$

where H_i = # Hispanics, i'th census tract;
B_i = # blacks, i'th census tract;
H, B = corresponding sums $\sum_i H_i$, $\sum_i B_i$.

The dissimilarity index has been in use for over twenty years (e.g., Taeuber and Taeuber, 1965; Marshall and Jiobu, 1975; Simkus, 1978). It has a number of important and useful properties: it ranges in value from 0 (no segregation) to 1 (complete segregation); its value can be calculated independently of the overall composition of different groups in the aggregate as a whole (e.g., overall percent black); it is applicable to situations where two groups are not exhaustive of the population; its value is easy to calculate, and it has an easily interpretable verbal description as well as a graphical one.

Expected Value of D

Recently, however, both Cortese, Falk and Cohen (1976) and Winship (1977) have described a major drawback of the dissimilarity index. A principal use of segregation indices is to compare different aggregates of establishments or environments with respect to their relative degree of segregation. For example, are blacks and whites more segregated in the first two years of college or in the last two years; are black and white managerial-level personnel more segregated from one another into different places of work than are black and white machine operators?

In the calculation of the segregation of a given aggregate, a certain degree of segregation may be produced by "chance." That is, a random allocation of the available blacks and whites to the establishments in the aggregate would not produce exactly the same racial distribution at each establishment. Clearly, if the establishments are quite small, although blacks might constitute only about 10% of a given population, even a random distribution might produce units that are 50% black.

Cortese, Falk and Cohen, and later Winship, showed that the expected value of the dissimilarity index--that is, its value under a condition of randomly produced segregation alone--is not only dependent on the size of the units in the aggregate, but also is highly dependent on the group composition (proportion black) of the aggregate. In particular, the expected value of D is larger as the aggregate increases in homogeneity, and smaller as the two groups considered become more similar

in total numbers. Table 1, adapted from Cortese, Falk, and Cohen (1976), shows the large magnitude of the dependency of the expected value of D on "percent black" and on the sizes of the units comprising the universe.[1]

For example, industry A with a minority population of 30% and an average of 100 workers per establishment would have an expected value of its dissimilarity index, E[D], of only .09; on the other hand, industry B with only a 5% minority concentration and an average employee population of 50 would have an expected value of .26 to its dissimilarity index. If the actual dissimilarity index for industry A was .25 and that for industry B was .30, it would seem more appropriate to consider industry A the more segregated even though its dissimilarity index was lower.

The Coleman Segregation Index

There is an alternative measure of segregation to the dissimilarity index--one that has many of the advantages of that index including its 0 to 1 range, common-sense interpretations, and its ease of calculation --but which is not nearly so subject to misleading results because of the "unequal expected value" problem. This index has been independently discovered on a number of occasions, as early as 1947,[2] but most recently by Coleman, Kelly, and Moore (1976) in their studies of racial segregation of public schools. Although this index goes by no common designation, we refer to it as the Coleman segregation index because of its genesis for our own use. However, where Coleman designated it as "r", we designate it as "S" for "segregation."

S is the proportional underrepresentation of one group in the environment of the average member of another group. It indicates the difference between the actual cross-racial experience of a group and the cross-racial experience that would exist under perfect integration of the two groups; that is, if all establishments had the same distribution of the two groups. Thus, it is a measure of lack of exposure given availability.

For example, in 1975, the privately employed labor force in the U.S. was 10.7% black. However, in the average (non-Hispanic) white worker's place of employment, only 8.7% of the workers were black. Therefore, the degree of segregation according to the Coleman S index was:

$$S = \frac{10.7 - 8.7}{10.7} = 19 \text{ percent} = .19.$$

Algebraically, the average percent black for whites is given by the sum of the proportion of blacks in each establishment, weighted by the number of whites at that establishment, divided by the total number of whites. Thus:

$$P_{b|w} = \frac{\sum (w_i \times p_i)}{\sum w_i}$$

where w_i = # whites, i'th unit (establishment);
p_i = % blacks, i'th unit (establishment).

The segregation index, then, is:

$$S = \frac{P - P_{b|w}}{P} = 1 - \frac{P_{b|w}}{P}, \quad \text{where } P = \text{overall \% black.}$$

The S index is not restricted to the case where the two groups considered form an exhaustive set of the groups in the universe. That is, "percent black" can refer to the proportion of blacks among all people in the particular unit, including those who are neither white nor black.

However, in such situations, the S index between two non-exhaustive groups--for example, between Hispanics and blacks--will differ depending on whether it is calculated based on the percent "X" for the average "Y" or the percent "Y" for the average "X." However, the two S values in this case are usually reasonably close. For example, the employment segregation between black workers and Hispanic workers is either .05 or .06 depending on whether it is calculated from "percent black for Hispanics" or from "percent Hispanics for blacks."

In contrast to the dissimilarity index, there is no comparable graphical interpretation to the S Index. However, when the two groups considered form exhaustive categories (blacks and non-blacks, for example), S has several interesting properties. To begin with, S equals the "between-establishments" proportion of the variance in the dichotomous individual-level variable, race, measured over all persons in the aggregate; That is:

$$S = \frac{\text{between SS}}{\text{total SS}} = \frac{\sum_i n_i p_i^2 - NP^2}{NP(1 - P)},$$

where
N = # establishments;
P = overall % black;
n_i, p_i = corresponding values of the i establishments.

This relationship is derived in Appendix 1.

Secondly, when the two groups are exhaustive, there is an additional interpretation of S. Namely, for blacks and non-blacks, for instance, it is the difference between the average percent black for blacks and the average percent black for non-blacks. In other words, if the average racial environment for blacks is 40% black and the average for non-blacks in the same establishments is 10% black, then S = .40 - .10 = .30. The derivation of this relationship is shown in Appendix 2.

Thus, the Coleman Segregation Index has three interpretations: it is the proportional underrepresentation of one group in the environment of the other in comparison to a situation of perfect integration; and when dealing with exhaustive categories, it is both the "between-establishments" proportion of the variance in the dichotomous variable distinguishing individuals in the two groups, and it is the absolute difference between the within-group environment for one group and the cross-group environment for the other, expressed as proportions.

Definitions of "Equal" Segregation and the Dependency on P in Calculating S-Index Scores

Taeuber and Taeuber (1965) have objected to indices similar to the Coleman segregation index because the group (racial) composition of the universe (P) is explicitly used in the derivation of index scores. In contrast, the dissimilarity measure is calculated independently of P. They suggest that the dependency on P in the calculation of an index makes it possible for different index values to be associated with situations that are "equally segregated" but which vary in P.

It is true that two universes with identical Lorenz curves but different values of P would have identical dissimilarity scores while the Coleman S scores for these two universes would not be the same. (The closer P becomes to 50%, the higher the S value would be.)

However, this point merely moves the question of what constitutes "equal segregation" back a step. The Lorenz curve, it may be recalled, is based on the distribution within each racial category across units of analysis. For example, consider two hypothetical "cities" with a

total of three "census tracts" each (Table 2). In both cities, one-tenth of all whites live in one tract along with half of all blacks. In a second tract, 40% of all whites and 40% of all blacks reside; and in the third tract, the remaining whites (half of all whites) and the remaining blacks (10% of all blacks) live.

In this situation, regardless of the relative numbers of whites and blacks, the dissimilarity index would be the same (D=.40). In contrast, the S score would vary depending on the value of P. S would equal .37 where P=.50, for example, and would equal .22 where P=.09 (or where P=.91). However, although from the perspective of the distribution within racial categories across tracts these two situations are equivalent, it is not clear that one would normally consider the situations to be "equally segregated."

Table 2 shows the breakdown of blacks and whites by tracts in two hypothetical universes. In the first, where P=.09, the three tracts have racial compositions of 2%, 8%, and 50% black. In the universe where P=.50, the three tracts are 9%, 50%, and 91% black. Variations in the racial composition are clearly larger on an absolute scale in the latter universe, and the between-tract proportion of the total variability of "race" is larger in the latter situation as well, as shown by the larger S value for that universe. From this perspective, then, the two situations are not equivalent and one would not want a segregation index to produce equal values for them.

Thus, the invariance of the two indices of segregation, D and S, across situations of "equal degrees of segregation" depends totally

on the definition of equal segregation that is used. It is not clear that the definition of equal Lorenz-curve inequality is necessarily superior to one based on explicit variability in group composition (e.g., percent black) of the different units.[3] In any event, the mere utilization of P in the calculation of the segregation index does not invalidate its utility or the appropriateness of the definition on which it is based.

Perhaps instead, the indices should be compared on other grounds.

Expected Value of S

We saw earlier that a major problem with the dissimilarity index was that its expected value differed markedly according to the number of people in each unit and according to the relative overall heterogeneity in the universe. How, then, does the expected value of S vary as a function of these two parameters?

In Appendix 3, we derive an approximation[4] for the expected value of S, $E[S] \cong k/N$, where k is the number of units (establishments) in the universe and N is the total number of persons in the universe. This solution applies to universes whose units vary in size; but in addition, where all units are of equal size n, $n = N/k$, and so E[S] simplifies to 1/n. Thus, where schools in a district are all size 500, or where they average size 500, $E[S] \cong .002$.

Two conclusions are apparent. First, the expected value of S, in contrast to that for the dissimilarity index, is not a function of P, the proportion of the universe population in group X. Secondly, the

expected value is much smaller relative to the range of possible values than is the case for E[D] and rapidly approaches zero for even moderately-sized average unit sizes. For example, where n = 25, $E[S] \cong .04$, whereas E[D] ranges between .16 and .79 (for $.01 \leq p \leq .99$). For n = 100, $E[S] \cong .01$; E[D] varies between .08 and .37.

Empirical examples

Although the dissimilarity index and the segregation index are derived from different definitions of segregation, they are measures of roughly similar concepts. Given appropriate conditions, their correlation across different aggregates should be substantial. For example, Zoloth (1976) found that the respective measures of between-school racial segregation of students correlated .87 across 2,393 school districts with at least a 5% minority student population.

However, there are situations where D and S have been found to be negatively correlated with one another. For example, in the same study referred to above, Zoloth found that the two indices of segregation used to measure between-school faculty racial segregation correlated -.24 across these 2,393 aggregates.

The difference between these two results is probably due to two factors: the significantly smaller size of the faculty populations at each school in comparison to the student populations; and the fact that the school districts included in the study were subject to a restriction that at least 5% of their student populations be from minority groups while there was no comparable restriction regarding the faculty ethnic composition. Thus the district-wide faculty varied over a wider range of "percent minority."

In Table 3, we present data from another situation where the dissimilarity index and the Coleman S index are not highly correlated. The data are from a study of racial segregation across places of employment for workers in the same general occupational category such as manager, clerical worker, or operative. The study covers a representative sample of 7,483 separate establishments in the private sector, each of which employs at least 25 workers (Becker, 1978).

Measures of racial segregation for the nine separate occupational categories are shown in the table. The dissimilarity index scores vary in the narrow range from .48 to .64. The Coleman S statistic varies between .14 and .40. (The fact that the S scores are nearly always lower is not significant because S is based on proportions of sums-of-squares, while D is based on proportions of cases.)

However, across these nine universes (occupational categories), the overall correlation between D and S is only +.26. The two categories with the lowest segregation measure by the Coleman S standard--managers and professionals--have two of the highest dissimilarity index scores. These are also the two occupational categories with the lowest percent black. As we indicated earlier, where there is a small proportion in the "minority" group, the expected value of the dissimilarity index is particularly large (see Table 1). Thus, the expected value of D for professional workers is approximately .45 while the expected value of D for laborers, for example, is only about .15. If one were to subtract the expected value of D from the actual scores (or alternately use a z-score transformation approximation suggested by Cortese,

Falk and Cohen), the resulting ordering of segregation by occupational category would be similar to but not identical with the ordering obtained from the Coleman statistic.

Applying a correction factor for the expected value to the Coleman S index (i.e. $S' = S - k/N$) does not change the relative ordering of the occupational categories. However, it does increase the distance between them: managers and professionals become even more distinctively the categories with the least racial segregation and segregation among operatives becomes more similar to the other lesser-skilled blue collar categories which have the highest degrees of segregation--laborers and service workers.[5]

It is instructive to note that the Coleman index values are highly correlated with the percent black in the occupational category--the greater the proportion of blacks in a category, the higher the segregation index value. We also found this relationship to be true in measuring the racial segregation of students across institutions of higher education. When institutions are clustered by state, the higher the black proportion of the college population in a given state, the higher the racial segregation among institutions within that state. (See Table 4; see also, Thomas, et al., 1976.)

We have seen, though, that the expected value of the Coleman segregation index is independent of the racial composition of the universe, and so these relationships must be produced by forces other than chance. It would be interesting to speculate about the causes of this association between percent black and the segregation index S, but we leave

that for another paper. Here we only wish to observe that the index that appears to be contaminated by the relative group proportions in the population is in fact free of such bias (i.e., the Segregation Index), while the index which appears empirically to be less correlated with the "percent black" factor and which is calculated independently of it (the Dissimilarity Index), is in fact the one whose expected value is strongly affected by this variable.

Decomposition of Segregation Indices--Controlling on Other Variables

One important property of a segregation index, as we have seen, is the resistance of its expected value to irrelevant variations among the aggregates whose internal segregation is being compared (such as average unit size and overall group heterogeneity). A second important property of a segregation index is its amenability to the decomposition of segregation into that portion that is between sub-classes of units within the aggregate and that portion that remains after controlling on the variable forming these sub-classes.

For example, how much of the observed black-white residential segregation is attributable to income differences between the races and how much to segregation within income classes? Or, how much of the racial segregation in employment is due to unequal distribution of black and white workers across the various occupational categories and how much is due to place-of-work segregation within occupational categories?

Winsborough (1974) showed how the dissimilarity index can be decomposed into "within" and "between" components. However, in this

decomposition there is a third component that cannot be allocated to either "within-compositional-category" segregation or "between-compositional-category" segregation. He describes it as "the differences in composition evaluated over the differences in composition-specific distributions" (p. 3).

Coleman's segregation index can also be decomposed when the two groups under consideration form exhaustive categories. Decomposition of the segregation index can be seen as an application of multi-level hierarchical analysis of variance. In the two-level version of this model, we have, for example, students within schools, schools within states, and states within a single region. The total sums of squares (equal to N*P*(1-P)) can be divided into portions representing variations within schools, between schools of the same state, and between states within the region. The proportion of the total sums of squares that is "between-schools/within-states" can be interpreted as the degree of segregation due to the varying group composition of the schools within the same state. The proportion of the total sums of squares that is "between-states" indicates the segregation due to variation in the group composition across states. Table 5 presents the formulas from the two-level hierarchical ANOVA model applied to the measurement of segregation.

Applying this method to the data on racial segregation of workers by place of employment and by occupational category (Table 3), we find the following results:[6]

A. Racial segregation by place of work for workers of all occupation categories combined..... .190 (from Table 3)

B. Racial segregation into different occupational categories and different places of work..... .301

(1) Segregation between occupational categories..... .047

(2) Segregation between places of work within occupational categories..... .254

In this example, we are using each worker's occupational category as the variable to split up the universe of establishments into different "between-" and "within-" groups. Although this seems like a different situation than the example of students-within-schools, schools-within-states, states-within-a-region, the principle is identical.[7]

Result "A" is reprinted from line 1 of Table 3; it refers to segregation measured when all workers at each place of employment are combined. In result B, we measure the segregation of workers both by place of employment and by occupational category. The level of segregation in result B is higher than that for result A. Part of this increase is due to the different racial proportions in the various occupational categories. Thus, result B(1) indicates that there is some racial segregation in employment across occupational categories. This is the segregation we are most aware of: i.e., blacks concentrated in service and blue collar categories. However, by far the greatest segregation in employment is across the various places of business within each occupational category (result B[2]). This measure of segregation is even higher than that given in A; that is, the racial

segregation among workers in particular occupational categories is higher than the overall racial segregation of workers as a whole.

Discussion and Conclusion

In this paper we have examined some of the characteristics of two statistical measures of the concept "segregation" as it has been applied to aggregates of collective units such as school districts containing schools, cities containing census tracts, and labor markets containing places of employment.

One of the two measures, the index of dissimilarity, or D, has been more widely used in social scientific analysis of segregation than any other index. This index, however, has been previously shown to possess a major disadvantage for the comparative study of segregation across different aggregates--namely that its expected value varies according to the size of the units in the aggregate and according to the group (racial) composition in the aggregate as a whole. In comparisons where aggregates vary sharply in racial composition and where the units tend to be small--for example, measuring segregation of different categories of workers across places of employment--this characteristic of the dissimilarity index can generate quite misleading results concerning the relative segregation of different aggregates. Of course, where these conditions do not apply--for example, when comparing the segregation of cities with similar racial composition using large census tracts as units of analysis--this attribute poses much less of a problem. However, other problems with the dissimilarity index--for example, its insensitivity to differences in segregation among the "over-represented" units or among those "under-represented"

by a group under discussion--may still prevent this index of derivations of it from being the index of choice (Winship, 1978).

In contrast, the alternative segregation index that we have discussed in this paper--the index referred to as the Coleman Segregation Index, or S--seems to have few of these drawbacks.

To summarize, S is the weighted average of cross-racial experience relative to the racial composition of the aggregate as a whole. It has several other verbal interpretations:

(1) S is the proportional underrepresentation of one group in the environment of the average member of another group;

(2) S is the between-establishments proportion of the variance in the dichotomous individual-level variable, race, measured over all persons in the aggregate; and

(3) When two groups are exhaustive, it is the difference between the average within-race experience of one group and the average cross-race experience of the other group.

In contrast to the often-substantial expected value of the dissimilarity index when people are randomly assigned to different units, the expected value of S approaches zero under nearly all circumstances. Also, it is independent of the racial proportions in the aggregate, whereas E[D] is not.

The fact that E[D] and E[S] behave so differently can account for a number of instances where D and S produce values that are either negatively correlated or only slightly positively related. Following up one of these instances, it was suggested that the observed association of S with the proportion black in an aggregate may be indicative of a

relationship of substantive importance--namely, that segregation is greater in universes with larger black proportions even when the measurement is standardized for racial composition.

Finally, it was shown that S can be decomposed into segregation between sub-classes of establishments in an aggregate and segregation between establishments within each sub-class. Racial segregation in employment provided one empirical example: segregation of workers of the same occupational category into different places of work is of much greater magnitude than the often-observed racial segregation that exists between occupational categories.

Thus, although the S index has not had as wide a usage as has the dissimilarity index, it does appear to be comparable or superior to the dissimilarity index in its interpretability, its ease of measurement, its decomposability, and especially in its robustness for inter-aggregate comparisons. Future research in this area should consider the segregation index used recently by Coleman, Kelly, and Moore, especially when comparing the relative segregation of aggregates that vary greatly in their overall group composition and where sizes of units under study are quite small.

Notes

1. Taeuber and Taeuber (1976), commenting on the Cortese, Falk and Cohen paper, pointed out that much of the reason for the high expected value of the dissimilarity measure under conditions of low percent minority has to do with the minimum value that the dissimilarity index can reach given the few persons that must be spread around the large number of units. Under these conditions, perfect integration is not possible. However, this fact illustrates the difficulty of using the index to compare the relative segregation of two different universes whose initial conditions vary so greatly. What is necessary for comparisons is an index whose expected values are invariant over such situations. See also Massey (1978) and Cortese, Falk, and Cohen (1978).

2. Duncan and Duncan (1955) refer to several discoveries including its discussion in Jahn, et al (1947) and Bell (1954). They refer to the index as the "eta" index while Taeuber and Taeuber (1965) discuss it as the Bell index.

3. Winship (1977), comparing D with two other measures--D - E[D] and an index identical to S--suggests that the comparison point of randomly produced segregation should be used when examining segregation as a dependent variable, but that study of the "effects of segregation" should employ D since "it makes little difference whether segregation is random or non-random." This is generally reasonable, except that the effects for individuals of the experience of racial isolation (Winship's example), are perhaps better measured by the individual's own specific racial experiences, and for the group as a whole, by "percent X for the average Y."

4. The need for the approximation symbol derives from the fact that in the theoretical model one is sampling units without replacement whereas the calculation of the approximate expected value assumes independent samples with replacement. Falk, Cortese, and Cohen (1978) have apparently found what they feel to be major discrepancies between the hypergeometric model (sampling without replacement) and the binomial model (sampling with replacement) for their analysis of E[D], particularly when the n_i vary in magnitude. On the basis of a reading of their published results, however, the discrepancy seems not to be sizeable; in any event, it is not clear that the same problem exists regarding the S index.

5. Another correction procedure, the use of a standard score $S' = \frac{S - E[S]}{\sigma^2[S]}$, would be superior to S. This is the procedure employed by Cortese, Falk, and Cohen (1976) to try to save the dissimilarity measure. Unfortunately, we have not obtained a solution to the value $\sigma^2[S]$ at this point, although Monte Carlo methods suggest themselves as a first step.

6. In contrast to the calculations for Table 3, where the denominator includes blacks, non-Hispanic whites and all others, these calculations consider only black and non-Hispanic white workers to preserve the exhaustive dichotomy required for the ANOVA interpretation.

7. In particular, we have workers of a given occupation (cf, students) working among others of the same occupational category at a place of work (cf, schools), which is one of the universe of such "work

groups" of similarly employed workers (cf, states), which, when combined with the other "work groups" of the other occupational categories, produces the entire labor market (aggregate) under study (cf, region).

References

BECKER, H. J. (1978) "Racial segregation among places of employment." Paper presented at the annual meetings of the American Sociological Association. (September)

BELL, Wendell (1954) "A probability model for the measurement of ecological segregation." Social Forces 32(May): 357-366.

COLEMAN, James S., Sara D. KELLY and John A. MOORE. (1975) Trends in School Segregation, 1968-73. Washington, D.C.: The Urban Institute.

CORTESE, Charles F., R. Frank FALK and Jack K. COHEN. (1976) "Further considerations on the methodological analysis of segregation indices." American Sociological Review 41 (August): 630-637.

--- (1978) "understanding the standardized index of dissimilarity: Reply to Massey." American Sociological Review 43 (August): 590-592.

DUNCAN, Otis Dudley and Beverly DUNCAN. (1955) "A methodological analysis of segregation indexes." American Sociological Review 20 (April): 210-217.

FALK, R. Frank, Charles F. CORTESE, and Jack COHEN. (1978) "Utilizing standardized indices of residential segregation: A comment on Winship." Social Forces (December).

JOHN, Julius, Calvin F. SCHMID, and Clarence SCHRAG. (1947) "The measurement of ecological segregation." American Sociological Review 12 (June): 293-303.

MARSHALL, Harvey and Robert JIOBU. (1975) "Residential segregation in United States cities: A causal analysis." Social Forces 53 (March): 449-460.

MASSEY, Douglas S. (1978) "On the measurement of segregation as a random variable." American Sociological Review 43 (August): 589-590.

MCPARTLAND, James. (1978) "Desegregation and equity in higher education and employment: Is progress related to the desegregation of elementary and secondary schools?" Law and Contemporary Problems.

SIMKUS, Albert A. (1978) "Residential segregation by occupation and race." American Sociological Review 43 (February): 81-92.

TAEUBER, Karl E. and Alma F. TAEUBER. (1965) Negroes in Cities: Residential Segregation and Neighborhood Change. New York: Atheneum.

--- (1976) "A practitioner's perspective on the index of dissimilarity." American Sociological Review 41: 884-9.

THOMAS, Gail E., Denise C. DAIGER, and James MCPARTLAND. (1978) "Desegregation and enrollment: Access in higher education." Johns Hopkins University. Center for Social Organization of Schools. Report.

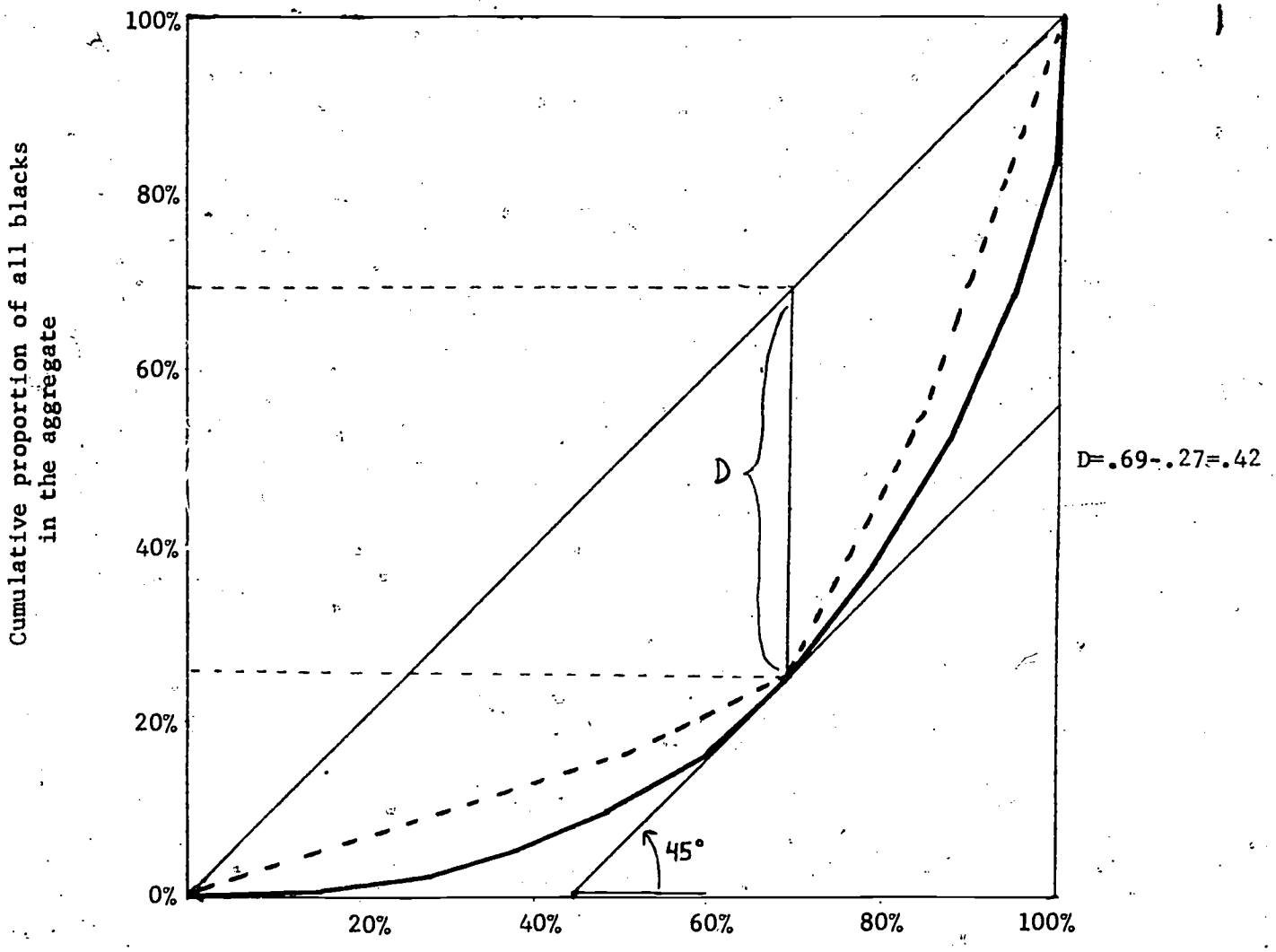
WINSBOROUGH, Hal H. (1974) "A note on the decomposition of indexes of dissimilarity." Discussion Paper 201-74. Institute for Research on Poverty. University of Wisconsin--Madison. (May)

WINSHIP, Christopher. (1977) "A reevaluation of indexes of residential segregation." Social Forces 55: 1058-1066.

-- (1978) "The desirability of using the index of dissimilarity or any adjustment of it for measuring segregation: reply to Falk, Cortese, and Cohen." Social Forces 57(December):717-720.

ZOLOTH, Barbara. (1976) "Alternative measures of school segregation." Land Economics 52 (August): 278-298.

Figure 1: Lorenz Curve and The Dissimilarity Index



*Units in aggregate ordered by "% black" in unit.

Table 1: Expected Value of the Dissimilarity Index under Random Distribution of "Majority" and "Minority" Populations across the Units in the Universe (exhaustive groups)

Expected Value of D:

Where the "minority's" proportion in the universe is . . .	And the number of "minority" plus "majority" people in each unit in the universe is . . . 10	25	50	100	1000
.01	.91	.79	.61	.37	.13
.02	.83	.62	.37	.23	.09
.05	.63	.37	.26	.18	.04
.10	.39	.27	.18	.13	.04
.20	.30	.20	.14	.10	.03
.30	.27	.18	.12	.09	.03
.50	.25	.16	.11	.08	.02

Adapted from: Charles F. Cortese, R. Frank Falk, and Jack K. Cohen, "Further Considerations on the Methodological Analysis of Segregation Indices," American Sociological Review. 41 (August, 1976): 630-637.

Table 2: Values of S and D under Two Hypothetical Situations

Situation I: Percent black = 50%

	Percent of all Blacks	Percent of all Whites	Number of Blacks	Number of Whites	Percent Black
Tract A	50%	10%	250	25	91%
Tract B	40%	40%	225	225	50%
Tract C	10%	50%	25	250	9%
Total, City I	100%	100%	500	500	50%

Segregation Indexes

D = .40

S = .37

Percent black for average white = 31.6%

Situation II: Percent black = 9%

	Percent of all Blacks	Percent of all Whites	Number of Blacks	Number of Whites	Percent Black
Tract D	50%	10%	50	50	50%
Tract E	40%	40%	40	450	8%
Tract F	10%	50%	10	500	2%
Total, City II	100%	100%	100	1000	9%

Segregation Indexes

D = .40

S = .22

Percent black for average white = 7.1%

Table 3: Segregation of Non-Hispanic Whites and Blacks across Places of Employment in the Private Sector, by Occupational Category

	Coleman Segregation Index	Dissimilarity Index	Percent Black	Mean Number Workers in this Category per Establishment*
All employees	.19	.48	10.7%	201
Managers	.15	.64	3.1%	22
Professionals	.14	.59	3.0	33
Technical Workers	.20	.57	6.5	25
Sales Workers	.21	.56	5.6	31
Clerical Workers	.20	.48	9.7	37
Craft Workers	.16	.49	7.1	45
Operatives	.23	.51	14.3	69
Laborers	.40	.62	21.0	41
Service Workers	.38	.58	22.5	32

* among establishments with one or more workers in this occupational category.

Source: Equal Employment Opportunities Commission. 1975. Annual Survey of Private Employers.

Table 4: Segregation of Whites and Blacks (others excluded) in 4-Year Colleges by State, Southern U.S. Region

	Coleman Segregation Index	Percent Black
Southern States	.60	17.3
Alabama	.65	24.9
Arkansas	.39	17.5
District of Columbia	.77	43.5
Delaware	.45	11.8
Florida	.47	12.6
Georgia	.64	22.8
Kentucky	.21	7.6
Louisiana	.57	25.9
Maryland	.52	19.4
Mississippi	.69	33.4
North Carolina	.74	21.2
Oklahoma	.25	6.5
South Carolina	.62	21.6
Tennessee	.49	14.6
Texas	.51	11.5
Virginia	.75	16.8
West Virginia	.09	5.5

Source: Office of Civil Rights, DHEW. 1976. Enrollments in Higher Education.

Table 5: Decomposition of the Coleman Segregation Index by Application of Hierarchical Analysis of Variance (2-level ANOVA example)

Source of Variation			Sums of Squares $X_{ijk} = [0, 1]$	Segregation Measure
General	Example			
Between Sub-aggregates	between states	= "mean-corrected between":	$\sum_i \left(\frac{X_{i..}^2}{N_i}\right) - \frac{(\sum\sum\sum X_{ijk})^2}{N}$	$\frac{\sum_i N_i P_i^2 - NP^2}{NP(1-P)}$
Between units within Sub-aggregates	between schools within states	= between schools - between states:	$\sum_i \sum_j \left(\frac{X_{ij.}^2}{n_{ij}}\right) - \sum_i \left(\frac{X_{i..}^2}{N_i}\right)$	$\frac{\sum\sum_{i\,j} n_{ij} p_{ij}^2 - \sum_i N_i P_i^2}{NP(1-P)}$
Between Individuals within units	within schools	= total - between schools:	$\sum_i \sum_j \sum_k X_{ijk}^2 - \sum_i \sum_j \left(\frac{X_{ij.}^2}{n_{ij}}\right)$	
Total	Total		$\sum_i \sum_j \sum_k X_{ijk}^2 - \frac{(\sum\sum\sum X_{ijk})^2}{N}$	

Where $N = \sum_i \sum_j n_{ij}$ $\quad N_i = \sum_j n_{ij}$ $\quad X_{i..} = \sum_j \sum_k X_{ijk}$ $\quad X_{ij.} = \sum_k X_{ijk}$

Appendix 1: S as Between-Establishments Proportion of Sums-of-Squares

We have that $S = 1 - \frac{\sum n_i p_i (1 - p_i)}{P(\sum n_i (1 - p_i)} = 1 - \frac{\sum n_i p_i - \sum n_i p_i^2}{P(\sum n_i (1 - p_i)}$

But $\sum n_i p_i = NP$ and $\sum n_i (1 - p_i) = N(1 - P)$

So

$$S = 1 - \frac{NP - \sum n_i p_i^2}{P \cdot N \cdot (1 - P)}$$

$$S = \frac{NP(1 - P) - (NP - \sum n_i p_i^2)}{N \cdot P \ (1 - P)}$$

$$S = \frac{NP - NP^2 - NP + \sum n_i p_i^2}{NP \ (1 - P)}$$

$$S = \frac{\sum n_i p_i^2 - NP^2}{NP \ (1 - P)}$$

From the analysis of variance of $X_{ij} = \begin{cases} 1 \text{ iff black} \\ 0 \text{ iff non-black} \end{cases}$

Sum of squares due to mean = $C = \frac{(\sum_i \sum_j x_{ij})^2}{N} = NP^2$

Between-establishment Sum Squares = $\sum_i \frac{(\sum_j x_{ij})^2}{n_i} - C = \sum_i n_i p_i^2 - NP^2$

Total sum of squares = $\sum_i \sum_j x_{ij}^2 - C = NP - NP^2 = NP(1 - P)$

(since $x_{ij}^2 = x_{ij}$ for $x = [0, 1]$)

$$\therefore \frac{\text{between SS}}{\text{total SS}} = \frac{\sum_i n_i p_i^2 - NP^2}{N P (1 - P)} = S \text{ (q.e.d.)}$$

Appendix 2: S as Difference between $P_{B|B}$ and $P_{B|\bar{B}}$, The Within- and Cross-Group Environments

Let $q_i = 1 - p_i$, the proportion of non-blacks in unit i, and $Q = 1 - P$

Then $P_{B|B} = \frac{\sum B_i P_i}{\sum B_i} = \frac{\sum n_i p_i p_i}{\sum n_i p_i}$ and $P_{B|\bar{B}} = \frac{\sum B_i P_i}{\sum \bar{B}_i} = \frac{\sum n_i p_i q_i}{\sum n_i q_i}$

$$P_{B|B} - P_{B|\bar{B}} = \frac{\sum n_i p_i^2}{NP} - \frac{\sum n_i p_i q_i}{NQ}$$

$$= \frac{Q(\sum n_i p_i^2) - P\sum n_i p_i(1 - p_i)}{NPQ} = \frac{Q(\sum n_i p_i^2) - P(\sum n_i p_i) + P(\sum n_i p_i^2)}{NPQ}$$

$$= \frac{P + Q\,(\sum n_i p_i^2) - P(\sum n_i p_i)}{NPQ}$$

$$= \frac{1 \cdot (\sum n_i p_i^2) - P \cdot NP}{NPQ}$$

$$= \frac{\sum n_i p_i^2 - NP^2}{NP(1 - P)}$$

$$= \frac{\text{between-establishment SS}}{\text{total sum-of-squares}} = S \quad \text{(q.e.d.)}$$

Appendix 3: E[S], The Expected Value of S, Equals $\frac{k}{N}$, The Average Number of Persons per Unit or Establishment. (Using a Binomial Approximation--Sampling with Replacement-- to the True Hypergeometric Distribution)

$$E[S] = E\left[1 - \frac{NP - \sum_{i=1}^{k} n_i p_i^2}{NP(1 - P)}\right] \qquad \left(N = \sum_{i=1}^{k} n_i\right)$$

$$= 1 - \frac{1}{NP(1 - P)} \left(NP - E\left[\Sigma n_i p_i^2\right]\right)$$

$$= 1 - \frac{1}{NP(1 - P)} \left(NP - \Sigma E[n_i] \cdot E[p_i^2]\right) \quad \text{(Since } n_i \text{ and } p_i \text{ are independently distributed.)}$$

Also, $\text{var } p_i = \frac{P(1 - P)}{n_i} = E[p_i - P]^2 = E[p_i^2] - 2 \cdot p \cdot E[p_i] + P^2 = E[p_i^2] - P^2$

$\therefore E[p_i^2] = \frac{P(1 - P)}{n_i} + P^2.$

Then $E[S] = 1 - \frac{1}{NP(1 - P)} \left(NP - \sum_{i=1}^{k} n_i \left(\frac{P(1 - P)}{n_i} + P^2\right)\right)$

$$= 1 - \frac{1}{NP(1 - P)} \left(NP - k\,P(1 - P) - NP^2\right)$$

$$= \frac{NP(1 - P) - NP(1 - P) + k\,P(1 - P)}{NP(1 - P)}$$

$$= \frac{k}{N}$$