DOCUMENT RESUME

ED 200 511 SO 013 283

AUTHOR McDonald, Scott C.
TITLE School System Response to Planned Interventions to Reduce Sex Bias.
PUB DATE 14 Apr 81
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Los Angeles, CA, April 14, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Administrators; *Educational Assessment; *Educational Needs; Educational Objectives; *Educational Practices; Educational Research; Elementary Secondary Education; Power Structure; *Sex Bias; Statistical Analysis
IDENTIFIERS Title IX Education Amendments 1972

ABSTRACT

A study was undertaken to discover the impact of strategies devised by the California Coalition for Sex Equity (CCSE) on reducing sex bias in schools. The strategies (referred to as power-based strategies) were based on identifying key decision makers in school districts and enlisting their support in helping staff identify and reduce gender bias in education and, specifically, in modifying institutional practices to conform to Title IX. The method involved using interviews and a pre- and posttest design to evaluate change among experimental and control groups in response to exposure and/or lack of exposure to the sex equity strategies. The sample consisted of administrators, teachers, students, and school board members in 23 experimental and 13 control school districts in California. Interviews and test scores were statistically analyzed. Findings indicated that school change processes are disorderly; network strategies, such as those which focused on interaction between teachers and administrators, are particularly effective in implementing affective and behavioral objectives; districts reach a threshold beyond which additional pro-equity training and services result in diminishing returns; and the power-based strategies were effective in combatting sex bias in all 23 school districts, although, of course, there were differences in degrees of effectiveness among school districts. The conclusion is that all school districts can benefit from power-based strategies to reduce sex bias and, in particular, those districts will benefit most which designate the superintendent or assistant superintendent to be the CCSE liaison, in which the teaching staff exhibits good overall morale, and in which flexiblity is stressed over bureaucracy and red tape. (DB)

SCHOOL SYSTEM RESPONSE TO PLANNED INTERVENTIONS TO REDUCE SEX BIAS

Paper presented at the annual meeting of the
American Educational Research Association
in Los Angeles, April 14, 1981

Scott C. McDonald
Department of Sociology
Harvard University

SCHOOL SYSTEM RESPONSE TO PLANNED INTERVENTIONS TO REDUCE SEX BIAS

Research Questions

In 1978 the California Coalition for Sex Equity in Education (CCSEE) received a grant under the Women's Educational Equity Act to study the effectiveness of its "power-based" strategies to reduce sex bias in public education. These strategies sought to be practical tools for identifying the key decision-makers in school districts, enlisting their support, modifying their institutional practices to conform to Title IX, and winning the "hearts and minds" of their staffs in the effort to reduce gender bias in education. The research grew out of practical concerns garnered in more than three years of work with school districts in California: What kinds of strategies were most effective? What school characteristics influence its propensity to change or to resist change? What features of daily educational practice were most malleable and which were most intractible? Most important of all, what was the net effect of the various workshops, seminars, and technical assistance provided by the member agencies of the California Coalition for Sex Equity in Education? Were districts any different as a result of their efforts?

It often is easier to pose research questions than to answer them. CCSEE immediately recognized one formidable barrier to the solution of its research problems: No measure of institutional sex bias existed. Without a measure, how could one discern change? How could districts be compared to each other or to themselves at different points in time? In order to compare districts on their Title IX compliance, some common metric was necessary--- some measure for "scoring" districts on their level of compliance. The ability of CCSEE to answer its practical questions about the effectiveness of its "power-

based" strategies hinged on its ability to measure changes in school district compliance with Title IX; hence, CCSEE endeavored to build and validate a measure of institutional gender bias.

Even if one were able to discern change in school districts using a theoretically and empirically perfect measurment instrument, the demons of skepticism are not easily exorcized. One must anticipate the possibility that measured changes are not caused by the power-based strategies per se, but by some hidden factors, some prior characteristics of the districts. While this possibility can be limited effectively by sampling strategies, it is nevertheless interesting and worthwhile to try to collect data on those prior characteristics that might interact with acceptance of Title IX. Hence, CCSEE also attempted to collect data on the ecological, organizational, fiscal, legal, stylistic, climatic, and "special" factors that accelerate or impede a district's willingness or ability to incorporate Title IX's imperatives into its educational practice.

Design of Research

The details of our research strategies were spelled out in earlier papers at AERA (McDonald, 1979), so I won't belabor the minutiae this morning. In brief, we employed a quasi-experimental pre-post research design. A stratified random sample of California school districts was selected to participate in the study, and assigned either to control or to experimental groups. Experimental group districts received the services of the CCSEE's power-based intervention strategies; control group districs did not. Pre- and post-treatment measures of Title IX compliance were taken, and some effort was made to record the districts' statuses on exogenous variables that

might affect receptivity to change. The measure of Title IX compliance itself was a detailed interview guide, complete with probing questions, covering all of the areas in which districts are, by law, supposed to investigate and eradicate gender bias. The interview procedure called for on-site interviews with teams of district personnel broadly representative of the various "actors" in the local educational process: administrators, union and non-union teachers, classified employees, students, and board members. Each interview question was accompanied by a Guttman scale depicting the ideal sequential steps that a conscientious district would take to address the equity questions raised by that particular question. These scales took the following general form:

A. District has taken no steps to address this point.

B. District has begun to investigate its behavior in this area by reviewing written documents, rules, policies, handbooks, etc.

C. District has further investigated its compliance in this area by collecting and analyzing quantitative data on patterns of participation, enrollment, employment, etc.

D. District has acted to remove inequities identified in steps "B" and "C" above.

E. Affirmative action is in evidence (i.e., District has removed barriers to equity and a pro-equity status-quo is in effect).

Hence, the Guttman scales assumed that districts behaved in a rational, sequential way. Recognizing that this might be a Pollyannaish view of school district processes, we hedged our bets by devising an alternative Likert scaling system. Under this system, each district was rated on a 5-point scale for its level of effort to comply with Title IX's requirements for each of the areas covered in the interview; these scale ratings permitted us to assign compliance "scores" to districts while simultaneously investigating their change processes through analysis of the Guttman scales. The interviews were scored by multiple raters to promote reliability, and a qualitative

validation of the results was sought through on-site visitations to a sub-sample of districts and through comparison to other diagnostic measures (where available).

The Sample

Using Department of Education data on per cent AFDC families in districts (arrayed in tripartite division) and on per cent minority enrollments (arrayed in quintiles), we selected our stratified random sample of school districts. California has an abundance of small, rural elementary school districts; to prevent them from distorting the representativeness of our sample, we allotted 75% of the sample slots to unified and union high school districts (found, for the most part, in the urban and suburban parts of the state). Even with a 100% oversample, however, we had difficulty filling the sample. We had the bad luck to be recruiting participants into our study during the summer of 1978, immediately following the passage of Proposition 13. Districts were extremely wary of involvement in any project, even if its promoters claimed that it would offer "free" services. The fiscal chaos and general battle-shock that beset districts during that summer made recruitment extraordinarily difficult. Although we had planned to have 30 experimental and 30 control districts, we only were able to enlist 23 experimental and 13 control districts. A comparison of the characteristics of California school districts and the districts in our sample appears in Table 1.

TABLE 1

Distribution of California School Districts According to Two Stratification Variables

	% Minorities					
% AFDC	1 (high)	2	3	4	5 (low)	N
1 (high)	136	77	55	43	38	349
2	39	96	85	61	63	344
3 (low)	26	32	68	100	123	349
N	201	205	208	204	224	1042

Districts Selected Into Sample, By Stratification Variables

		% Minority Enrollment						
% AFDC	Treatment	1 (high)	2	3	4	5 (low)	Total AFDC	N
1 (high)	Experimental	2	2	1	1	1	7	
	Control	1	0	2	1	0	4	
2	Experimental	1	4	3	1	0	9	
	Control	1	2	0	1	2	6	
3 (low)	Experimental	0	1	1	2	3	7	
	Control	0	1	0	1	1	3	
Total N Minority								
	Experimental	3	7	5	4	4		N=36
	Control	2	3	2	3	3		

Although our sample was generally representative of the ethnic and socio-economic diversity of California and though its randomness allowed it to be rid of obvious selection biases, its puny size presented statistical problems. One cannot safely generalize from small samples. Nearly all statistical techniques demand considerably larger samples, and small-samples usually never allow the investigator to control for exongenous variables. These problems were serious, but not fatal. A sample so nearly purely random, though small, can suggest the causal relations one would likely find in a larger sample. Though the size of the sample limited the statistical arsenal on which we could draw, we fought our analytic battles with all available weaponry. (Excuse the martial metaphor). Our key results follow.

Efficacy of the Measure of Institutional Sex Bias

The details of our validation procedures have been detailed elsewhere (McDonald, 1979; Mahon and Peterson, 1981), so I'll be brief here. Suffice to say that our procedures rested on three pillars: (1) assessment of inter-rater reliability of Likert-type ratings; (2) qualitative comparison of Likert-scale results to observations made in verification site visits (and, where available, by comparison to reports from other independent data collection efforts); (3) scalogram analysis of the Guttman scales. The first two procedures convinced us that, for the most part, the Likert scaling was accurate. I say "for the most part" because the scores for three districts (two experimental and one control) could not be reconciled with the qualitative data; as a result, these three cases were dropped form further analysis. For the remaining 33 cases, however, Likert scale scores agreed substantially with written reports on district status, with data from un-obtrusive observations made during verification site visits, and with other

independent sources of data on the districts. Furthermore, independent Likert-type ratings using the tape recorded interviews as common stimulus led to a high degree of agreement on ratings (r= .92). Hence, we took the Likert scale scores to be an accurate reflection of the Title IX compliance status of the 33 districts remaining in the sample.

If our Likert scaling procedure passed its tests with flying colors, our Guttman scales flunked miserably. By applying the stringent conditions of the Goodenough technique of scalogram analysis to the completed Guttman scale items, not one of our 40 interview-related scales displayed adequate sequential, cumulative structure. In other words, none of our Guttman-like items met our scalability criteria. Three general explanations for this are plausible.

(1) School change processes might themselves be disorderly and non-sequential.

(2) The wording of the scales might have erred by being too detailed (i.e., with more scale steps than are needed to capture district transitions). Alternatively, the scores might have mis-specified the actual compliance steps taken by the districts, or they might have mis-specified the sequential order in which those changes take place. Any of these scale mis-specifications could have led to the erratic scale-item correlations that our analyses obtained.

(3) Despite our elaborate precautions in the training of interviewers, our field staff might have been confused about the mechanics of the scales. In particular, some interviewers/raters might have failed to realize the importance of checking all applicable scale items--- not just the "most applicaple" or the "highest

applicable" items. This problem could have been exacerbated by the cumulative presumptions implicit in the wording of some scale steps (e.g. "Based on the steps taken in 'b' and 'c' above, district has. . ."). Any of these problems might have led interviewer/raters to mark fewer scale items than were, in fact, relevant---thereby yielding high modal responses that increase the minimum marginal reproducibility of the scales but, by implication, depress the coefficients of scalability.

Further research would be needed to disentangle the puzzles presented by our Guttman scale responses. In particular, it would be worthwhile to try a variety of scale wordings and formats in school districts, using similar qualitative validation procedures to assay the accuracy of the responses. This could be accomplished efficiently and simply without the elaborate pre-post random sample design employed in this study. As it stands, however, we cannot determine with any certainty whether our Guttman scaling attempt failed because of our scale construction and implementation, or because school district change processes <u>really are</u> as helter-skelter as our results would suggest.

<u>Did the Planned Interventions (or "Power-Based Strategies") Appreciably Help Districts Comply with Title IX?</u>

As I have already noted, the analyst working with small samples has an extremely limited repertoire of statistical techniques at his or her disposal. Even if the mean score differences between experimental and control groups were quite large, our sample was too small to perform a T-test of difference in means. Furthermore, our mean scores were swamped by their large standard deviations, and scatterplots of our univariate score distributions showed lots of exotic shapes, but no friendly bell-shaped curves. Hence, there was absolutely

no basis for assuming the normalcy of the underlying distribution (although a larger sample, indeed, may have shown normal tendencies). Our solution was to use the non-parametric Mann-Whitney U-test where actual scores are discarded in favor of the score rankings, thus providing a test that is not affected by skewness or any other distributional peculiarity (i.e., a distribution-free test). As such it is not distorted by extreme scores, and it has demonstrated high asymptotic relative efficiency (relative, that is, to the T-test), even when samples are small and populations are not normal (Wonnacott and Wonnacott, 1972). In effect, then, we tested for the differences in score rank between the experimental and control groups, prior to treatment and after treatment. We also tested for rank differences in gain scores, a more conservative test that standardizes for any initial score advantages enjoyed by districts.

We found that, at the outset, true to our sampling intentions, there were no significant score rank differences between the two groups. We did detect a slight (non-significant) difference in the area of athletics, but it was to the advantage of the experimental groups--- control group districts were slightly more compliant in this area than were experimental groups (see Table 2).

Table 2

Mann-Whitney U-Test for Differences Between Experimental and Control Groups, Pre-Treatment

Title IX Dimension	Mean Rank Control (N=12)	Mean Rank Experimental (N=21)	U	Z	2-Tailed P-Value
Access to Courses	16.96	17.02	125.5	-0.019	.985
Non-Academic Courses	16.87	17.07	124.5	-0.056	.955
Physical Education	17.12	16.93	124.5	-0.056	.955
Athletics	20.12	15.21	88.5	-1.404	.160
Employment	17.58	16.67	119.0	-0.262	.793
Minimal Compliance	15.83	17.67	112.0	-.526	.599
Total Score (All Dimension)	17.00	17.00	126.0	0.000	1.000

A year and a half later, after receiving training and technical assistance from CCSEE, the experimental group district had improved its score rankings to such an extent that statistically significant differences existed in the areas of "access to courses," "physical education," "employment," "minimal compliance," and overall "total score" on Title IX compliance. (See Table 3)

Table 3

Mann-Whitney U-Test for Differences Between Experimental and Control Groups, Post-Treatment

Title IX Dimension	Mean Rank Control (N=12)	Mean Rank Experimental (N=21)	U	Z	2-Tailed P-Value
Access to Courses	9.50	21.29	36.0	-3.369	.001*
Non-Academic Activites	12.83	19.38	76.0	-1.872	.061
Physical Education	11.79	19.98	63.5	-2.340	.019*
Athletics	14.50	18.43	96.0	-1.123	.261
Employment	12.33	19.67	70.0	-2.096	.036*
Minimal Compliance	12.21	19.74	68.5	-2.160	.031*
Total Score (All Dimensions)	9.83	21.10	40.0	-3.218	.001*

The difference between the two groups in the area of "non-academic activities" nearly attained the .05 criterion level of statistical significance, but fell slightly short. The experimental group districts did not overcome their initial disadvantages in athletic compliance enough to, at post-treatment, be significantly ahead of the control group districts in this area.

The best test of the effectiveness of the intervention startegies, however, is found not in a static time-freeze comparison of the two groups at any particular point in time, but in a comparison of their relative progress over time. Table 4 shows the results of a U-test of rank-order differences in gain scores (ie. the difference between pre-treatment and post-treatment scores) between the two groups.

Table 4

Mann-Whitney U-Test for Differences in Gain Scores
Between Experimental and Control Groups

Title IX Dimension	Mean Rank Control (N=12)	Mean Rank Experimental (N=21)	U	Z	2-Tailed P-Value
Access to Courses	10.75	20.57	51.0	-2.807	.005*
Non-Academic Courses	14.50	18.43	96.0	-1.123	.262
Physical Education	11.79	19.98	63.5	-2.339	.019*
Athletics	12.79	19.40	75.5	-1.890	.059
Employment	11.92	19.90	65.0	-2.283	.022*
Minimal Compliance	13.79	18.83	87.5	-1.451	.147
Total Score (All Dimensions)	10.83	20.52	52.0	-2.769	.006*

Here we see clearly that the intervention "treatments" led to significant gains in the areas of "access to courses," "physical education," and "employment," as well as in the overall sum of all dimensions ("total compliance"). We did not obtain sizable experimental group gains in the areas of "minimal compliance" and "non-academic activities". The absence of noteworthy gains in the former area probably stems from the fact that most districts had met nearly all of their minimal compliance requirements before becoming involved in the project--- hence they had little room in which to improve in that area. The lack of improvement in experimental group compliance in "non-academic activities" probably testifies to the difficulty of effecting (and measuring) change in this most amorphous area of school practice.

What Other Factors Affect District Response to Interventions?

Our use of the Mann-Whitney U-test, while statistically defensible, limited us to the dichotomous question of whether there was any difference between the two groups. To probe the more sophisticated questions, we resorted to a procedure for which no purely statistical case can be made. Since any attempt to examine the simultaneous effect of three variables on only 33 cases results in very small frequencies in table cells, we could make no grandiose claims that our sample justified statistical inference to the universe of school systems in California, much less the nation. Our data, however, did appear to be fairly good and our sample unbiased. Although our small sample size made our analyses extremely vulnerable to sampling error, the presumed absence of selection bias suggested our data's ability to whisper real causal relations to the attentive ear. So we took the leap of statistical faith and employed a variant of D-systems analysis, a procedure for the analysis of categorical data by which one can make statistical inferences with proportions and model causal relations among categorical variables (Davis, 1975; Davis, 1976; Goodman and Magidison, 1978). The confidence intervals in D-systems are sensitive both to the extremeness of the proportion differences and to the marginal frequencies (ie. the marginal sample sizes). Our small sample size, taken at face value, would have led to confidence intervals that would have swamped even the most extreme differences in proportions. However by making the very optimistic assumption that the sample was not biased and that the addition of more cases would have yielded, more of less, the same results, we were partially freed from the paralysis of our small sample size. In short, though we analyzed actual proportions and marginal frequencies, we computed confidence intervals based on a fictitious amplification of our table cells by a factor of ten. Using this artifice, some differences in proportions

became salient enough to protrude beyond their confidence intervals; only D's that met this arbitrary criterion were modeled in our linear flow graphs. Variances were computed on the assumptions of a simple random sample, using a sigma value of 1.96. Chi-square tests for the significance of table interactions were computed, based upon the same artificial enlargement of table cell frequencies, and significant interactions were designated on the linear flow graphs. Of course, the actual differences in proportions were not affected by the arbitrary inflation of the table cells; only the confidence intervals were made smaller. This procedure does not permit any statistical inference (as true D-systems analysis would); rather it served as a sorting device by which we might identify the most salient effects. There are, no doubt, distrotions in this procedure--- particularly since the empty cells that remain empty when multiplied by 10 would probably have at least a few cases in them in an actual sample of 330 school districts. Hopefully the quality of our small sample minimized the pernicious effects of these distortions. Having confessed this rather unconventional statistical solution to the problem of small-sample paralysis (and duly warned readers about its limitations), I proceed to our substantial findings.

Modified D-systems analysis allowed us to see whether there appeared to be any relationship between the specific services provided to districts and their specific areas of score gains in Title IX compliance. Indeed, there was a striking correspondence. Except in the areas of "non-academic activities" and "employment," districts gained in the particular dimensions of Title IX that their specific intervention services from CCSEE had addressed. I see this as further evidence of the effectiveness of intervention services provided by CCSEE.

A comparison of the efficacy of different service stategies was not particularly illuminating since no single strategy (e.g. diagnosis, legal pressure, consciousness-raising workshops, etc.) was associated with score gains. This suggests that all approaches are equally advantageous. The notable exception, however, was the "resource linkage/networking" strategy--- an approach that clearly emerged from the pack and demonstrated greater effectiveness.

Our data did not permit discernment of the functional relation between services and gains. That is to say, our data were much too thin to allow us to detect linearity, "threshold effects", or the like. However, most of the tabular relationships between services and gain scores were free from statistical interactions (ie. the direction of the effects was consistent). This at least whispers the possibility of some linear effects. However, our results also suggest that interventions may have "threshold effects". Those districts that had had considerable prior experience with sex equity projects, for the most part, did not register significant gains during their tenure in this study. This result suggests that, after initial exposure to pro-equity training and technical assistance, districts reach a threshold beyond which additional services are greeted by diminishing returns.

Our review of exogenous factors affecting district acceptance of Title IX was, of course, the aspect of the study most severely compromised by our small sample size. However, amidst all of the qualifications and conditions, a few cautious conclusions can be advanced. First and foremost, it appears that the treatment effects detected by the Mann-Whitney U-test survived virtually all controls introduced in the D-systems analysis. That is to say, the differences between the experimental and control groups were not due to any hidden or exogenous factors, nor to compositional differences between the groups. Furthermore, the D-systems analysis confirmed our hope that our sample

was unbiased in most substantively important respects.

Our D- systems analysis also permits us to draw a composite sketch of the "high impact" districts (ie. the districts that tended to make the greatest strides toward Title IX compliance). Demographically, they were:

---elementary school district
---smaller districts (in terms of number of shcools, number of employees, and average daily attendance)
---non-metropolitan districts
---districts that had not had any prior contact with pro-equity training and technical assistance programs.

When we considered the internal Title IX compliance structures of the "high impact" districts, we found that they were:

---districts that had designated the Superintendent or the Assistant Superi tendent to be the liaison to CCSEE
---districts in which the Title IX officer had flexible (ad hoc) time commitments to her of his Title IX duties.

The "high impact" districts also were:

---districts that had endured relatively little fiscal trauma as a result of Proposition 13's revenue reductions
---districts that were marked by flexibility rather than by cumbersome bureaucracy and red tape
---districts in which the teaching staff exhibited good overall morale.

A similar composite sketch can be drawn for those districts that declined in level of Title IX compliance (ie. had lower post-treatment compliance scores than they had at pre-treatment); these "decliner" district were:

---districts that serve more affluent neighborhoods
---districts that designated a person at the sub-cabinet level to serve as liaison to CCSEE
---districts that are burdened by cumbersome "red tape".

Finally, our D-systems analysis shows that the group of districts that neither improved nor worsened (ie. the "no change" group that remained virtually stationary during the two years of the study) were characterized by:

---considerable sex equity activity prior to CCSEE
---having had grievances filed during participation in the study.

While these findings do not deserve our endless confidence, they are strong enough and consistent enough to warrant our serious consideration and discussion. No study can provide results formidable enough to justify bland acceptance or termination of further questioning. Hopefully, this study will have the opposite effect--- the opening of new avenues of inquiry by both sex-equity researchers and practitioners. With diligence and a little luck, our efforts will, in the long run, be so enhanced that on each future occasion when a consultant walks into an inservice training meeting, a board room, or a playing field, the groundwork will have been laid for a successful endeavor. Armed with better knowledge about our audiences, we may hasten the arrival of that new morning in America when "equity" for all people is not a hollow promise, but a reality.

REFERENCES

Davis, James A. 1975. "Statistical Inference with Proportions." Mimeo. National Opinion Research Center, Chicago.

Davis, James A. 1976. "Analyzing Contingency Tables with Linear Flow Graphs: D-Systems," in David Heise, ed., Sociological Methodology. San Francisco: Jossey-Bass.

Davis, James A. 1978. "Contingency Table Analysis: Proportions and Flow Graphs." Mimeo. Harvard University.

Goodman, Leo and Magidson, Jay, eds. 1978. Analyzing Qualitative Categorical Data. Cambridge, Mass.: Abt Books.

McDonald, Scott C. 1979. "Methodological Dilemmas in the Assessment of School Interventions to Promote Sex Equity (and a Progress Report on One Experimental Resolution)." Paper presented at annual meeting of the American Educational Research Association in San Francisco. April 11, 1979.

Mahon, Lee and Peterson, Barbara. 1981. "School Change Processes: Can They Be Measured?" Paper presented at annual meeting of American Educational Research Association in Los Angeles, April 14, 1981.

Wonnacott, Thomas J. and Wonnacott, Ronald J. 1972. Introductory Statistics for Business and Economics. New York: John Wiley and Sons.