Journal of Consulting and Clinical Psychology
1990, Vol. 58, No. 6, 720–728

0022-006X/90/$00.75

Arrests Among Emotionally Disturbed Violent and Assaultive Individuals Following Minimal Versus Lengthy Intervention Through North Carolina's Willie M Program

John R. Weisz
University of North Carolina at Chapel Hill

Bernadette R. Walter
Orange-Person-Chatham Mental Health Center
Chapel Hill, North Carolina

Bahr Weiss
Vanderbilt University

Gustavo A. Fernandez and Victoria A. Mikow
Division of Mental Health, Developmental Disabilities, and Substance Abuse Services
North Carolina Department of Human Resources

Time to first arrest after termination of Willie M Program services was compared in 2 groups of former clients. All Ss had met program criteria and "aged out" after their 18th birthday, but the two groups differed in duration and extent of intervention received: (a) A short-certification group (n = 21), because they turned 18 near the 1981 program start date, had received Willie M services for a mean of only 26 days (all cases < 3 months); (b) a long-certification group (n = 147) averaged 896 days in the program (all cases > 1 year). The groups did not differ significantly in gender or race; geographic region; IQ; diagnosis according to the *Diagnostic and Statistical Manual of Mental Disorders,* 3rd. ed. (DSM-III, American Psychiatric Association, 1980); or age at earliest antisocial acts. A survival analysis compared the short and long groups on proportion avoiding arrest as a function of time since aging out. The long group showed slightly better arrest survival, but survival curves for the 2 groups did not differ reliably. Thus the program was not found to significantly reduce the risk of young adult arrests.

Accumulating evidence from outcome research suggests that mental health interventions for many types of emotionally disturbed children and adolescents can have beneficial effects (see meta-analyses by Casey & Berman, 1985; Weisz, Weiss, Alicke, & Klotz, 1987). However, the evidence on interventions for young people who show delinquent, antisocial, and violent behavior is more mixed. For example, in Davidson, Gottschalk, Gensheimer, and Mayer's (1984) meta-analysis of 91 intervention studies with delinquent youth, 60% of the studies yielded some evidence of positive intervention effects, but "when the actual size of the effects is considered, a radically different conclusion is reached. . . . Namely. . . there is no evidence that interventions with juvenile delinquents produce positive effects" (p. 28). This conclusion resembles earlier judgments (on both juvenile and adult intervention programs) by Martinson (1974) and by Greenberg (1977), who concluded, "The blanket assertion that 'nothing works' is an exaggeration, but not by very much" (p. 141).

In contrast, Garrett (1985) conducted a meta-analysis of 111 intervention studies with delinquents and concluded that her results "suggest that treatment of adjudicated delinquents in an institutional or community residential setting does 'work' " (p. 287). However, Garrett did note that the average improvement in delinquents as a function of treatment was only 0.37 of a standard deviation across various outcome measures. And when the analysis was confined to more rigorous studies (e.g., those using random or matched control groups), the figure fell to 0.24 (cf. 0.71 and 0.79 in the psychotherapy meta-analyses by Casey and Berman [1985] and Weisz et al. [1987]); this means that in the rigorous studies, after treatment, the average treated delinquent was at only the 59th percentile of the control group delinquents, averaging across measures. Of course, even effects of this magnitude, if reliable, might well be beneficial to society.

This study was made possible by funding to John R. Weisz through the Division of Mental Health, Mental Retardation and Substance Abuse Services, Department of Human Resources, State of North Carolina, and through grants from the National Institute of Mental Health. We are grateful for this support and for assistance in gathering information—assistance provided by the State Bureau of Investigation, Department of Corrections, and numerous individuals at local mental health centers.

We are grateful to Ken Parker of the North Carolina Department of Corrections for his help in obtaining arrest data. We appreciate the advice and suggestions of Denise Barnes, Charles Bullock, Robert Cairns, Lee Kittredge, David Langmeyer, and David Rogers, and we thank Joel B. Greenhouse for his very thoughtful review of the manuscript. We also thank Frank Bostancic, Janet Clarke, Jamie Crist, Susan Glick, Elizabeth Jackson, Kelly Jones, Mary Beth Kent, Michael Lambert, Todd Morton, Jim Piasecki, Val Proffitt, Lisa Richardson, Dan Sussman, Greg Tourian, and Aviva Wasserman for their help with various aspects of the study. Finally, we are grateful to the young adults who participated in this study.

Correspondence concerning this article should be addressed to John R. Weisz, who is now at the Department of Psychology, Franz Hall, University of California, Los Angeles; 405 Hilgard Avenue, Los Angeles, California 90024-1563.

Even in the most recent round of meta-analysis and interpretation, different writers reach rather different conclusions about what the data show. Whitehead and Lab (1989) examined 50 post-1974 studies of intervention programs designed to inhibit recidivism (deviant acts or further offending) among juvenile offenders. Summarizing their findings, Whitehead and Lab stated:

> The results show that interventions have little positive impact on recidivism and many appear to exacerbate the problem. . . . the earlier evaluations that claim that "nothing works" are close to the conclusion to be drawn from more recent evaluations of juvenile treatments. (p. 276)

Recent reports by Lipsey (1988, in press) are somewhat more optimistic. Lipsey's (1988) review of three major meta-analyses has led him to conclude that "even the typically low-grade research on delinquency treatment available in existing literature reveals positive treatment effects of modest, but not trivial magnitude" (p. 80). In Lipsey's (in press) own meta-analysis of 443 studies, he found effects that were positive and significant, though modest. Pooling across studies and delinquency measures, Lipsey found that treated juveniles averaged about .10 standard deviation units less delinquency after treatment than did control group juveniles. Of particular interest was the wide variability in effects of the various interventions, indicating that under some circumstances strong positive effects are found. "The challenge," noted Lipsey, "is to discover the nature of those circumstances" (p. 40).

One possible set of optimum "circumstances" has been proposed by Andrews et al. (in press), who reanalyzed the Whitehead and Lab (1989) data described above (combined with additional studies). Andrews et al. argued that what works best is "appropriate" service, which they have defined as delivery of intervention to high-risk cases, targeting "criminogenic" needs, and matching treatment approaches to the specific needs and learning styles of the clients. Studies of interventions that met the authors' criteria for appropriateness were reported to show a strong mean effect size, higher than 0.6 standard deviation units. Such a finding supports those who have argued that different types of treatment are differentially effective with different types of youth and that for optimum effectiveness we will ultimately need to match treatments to individual youths (see, e.g., Barrett, Hampe, & Miller, 1978).

This is the philosophy behind the program examined in the present study. The Willie M Program was developed in North Carolina in response to a class action suit on behalf of seriously emotionally, neurologically, or mentally handicapped youth who are violent or assaultive. Established in 1981, the program has served over 2,800 youth at a cost of more than $250 million. The core of the program is intensive case management. Each Willie M class member is assigned a trained case manager who works with a clinical/educational treatment team to determine the class member's treatment needs and arrange for an individually tailored array of services targeted to those needs. Officials classify the services into about 20 traditional categories (e.g., outpatient psychotherapy, inpatient treatment, recreation). To be certified as class members, individuals must be judged "seriously emotionally, neurologically, or mentally handicapped" and must show "violent or assaultive behavior" (North Carolina Department of Human Resources, 1989, p. 30). Once certified, youths receive services averaging about $25,000 per year. During the period studied here, young people "aged out" of the program 6 months after turning 18, or at the end of the same fiscal year, whichever period was longer. (For detailed descriptions of the program, certification, and the population served, see Behar [1985]; Curry, Pelissier, Woodford, & Lochman [1988]; and Macbeth [1985].)

It has been suggested that the program may serve as a model of appropriate intervention with seriously disturbed youngsters (Behar, 1985; Macbeth, 1985), and the program does show some of the characteristics of "appropriateness" described by Andrews et al. (in press). However, there has not yet been empirical assessment of program effects. Such assessment has been hampered by a key problem: the lack of a true no-treatment control group to which the treated youth can be compared. Not only are program officials reluctant, on ethical grounds, to assign any class member to a no-treatment control condition, but the consent decree associated with the program prohibits denial of services that are deemed needed and appropriate. Thus, forming a true randomly assigned no-treatment control group for purposes of outcome research is not merely difficult in a practical sense, but it is illegal.

Consequently, only *indirect* approaches to the question of program effects are possible, approaches that may depart from conventional methods involving randomly assigned treatment and control conditions. Although clearly less than ideal, such indirect approaches may be the only way to generate useful information. Under such circumstances, it will be important for investigators to recognize the limitations their methods impose on interpretation of findings.

In our study, we sought information bearing on effects of the Willie M Program. Given the constraints noted above, we were required to consider approaches that did not involve a randomly assigned no-treatment control group. We considered comparison groups from other states as well as comparison groups from North Carolina who had been nominated for the program but had failed to meet all certification criteria. Neither type of group could be considered truly comparable to the Willie M population. Ultimately we settled on a comparison of two highly similar groups, both comprising individuals fully certified as Willie M class members in the early years of the program. Members of one group all were certified for more than one year and had received substantial services. Members of the other group, primarily because their 18th birthdays happened to fall near the time the program began, had been certified so near their legally mandated "aging out" date that they had actually been served for 90 days or less. Because establishing a relationship with a case manager and processing applications for services are time-consuming processes, this group had received minimal service through the Willie M Program; indeed, a number of these had received little intervention beyond assignment of a case manager.

These short-certification (short-cert.) and long-certification (long-cert.) groups, which proved to be quite similar in demographic characteristics and problem behavior histories at the time of certification, were compared on one possible measure of program effects: time to first arrest after aging out. This is but one of several possible outcome indicators; other researchers might prefer to focus on more social or psychological outcomes. We focused on arrests because of their societal importance.

Method

Sample Selection Procedures and Resulting Sample

The full population of Willie M class members who aged out of the program between its inception in October 1981 and January 1986 contained 794 individuals. For our initial sample pool, we selected all 23 members of this population who had received services for 90 days or less prior to aging out (short-cert. pool). From the full population we also randomly selected an additional 200 who had received one year or more of services prior to aging out (long-cert. pool). From the short- and long-cert. groups thus selected, we dropped a total of 14 who were in prison or jail at the time of aging out (and thus unavailable for arrest), 8 who spent part of their risk period (i.e., the period from decertification to first postdecertification arrest) in prison or jail for offenses committed prior to aging out, 16 who had disappeared, 7 whose parents had refused Willie M services for their child, 9 who had themselves refused services (as emancipated minors), and 1 who had entered military service. These procedures resulted in a final sample of 168, 21 subjects in the short-cert. group and 147 in the long-cert. group. The short-cert. group averaged 26.24 days receiving services in the program; the long-cert. group averaged 895.9 days.

Willie M Treatment Program and Services Received by Each Group

In the Willie M Program, the case manager and treatment team work to fit services to the particular needs of the treated individual. Services are grouped by the program into approximately 20 categories. Although most of the categories (e.g., respite care for parental relief, recreation, after school programs) have at least some relevance to mental health, the most directly relevant appear to be (a) inpatient therapy, (b) outpatient child therapy, (c) outpatient family therapy, (d) day treatment, (e) supervised group living services, (f) supervised independent living services, and (g) vocational placement. To assess whether the short- and long-cert. groups were actually different in mental health services received, we compared the two groups on the number of months in which services in each of the 7 categories were received. We found significant differences in each category, as shown in Table 1.

Comparability of the Short- and Long-Cert. Groups

In general, whether a child was a member of the short- or long-cert. group depended on the year he or she was born in relation to when the Willie M Program began. Thus we had no a priori reason to suspect that the groups would differ along important clinical or demographic dimensions. (Of course, the groups did differ in age at certification, inasmuch as, by definition, all short-cert. cases were near age 18 when certified; see Table 1.) Nonetheless, it seemed important to establish the equivalency of the two groups, given that individuals were not randomly assigned to groups. Using chi-square and *t* tests, we compared the groups on several demographic and psychological characteristics. As shown in Table 1, our tests revealed that the groups did not differ in race; gender; region of residence within the state (i.e., West, North Central, South Central, or East mental health administrative district); full-scale Wechsler Intelligence Scale for Children—Revised (WISC-R) IQ; whether the subject had received at certification a *DSM-III* psychosis diagnosis (schizophrenia, paranoia, or atypical psychosis); or whether the subject had received a conduct disorder diagnosis. The last three variables were derived from diagnostic workups prepared by licensed psychiatrists and psychologists as part of the Willie M nomination and certification procedure.

A number of studies (reviewed by Loeber, 1988; Loeber & Stouthamer-Loeber, 1987) have suggested that one important predictor of later arrest is the age at which youngsters begin to show antisocial behavior. Consequently, we also compared the short- and long-cert. groups on the age at which the subjects reported that they had first committed various antisocial acts. The self-reports were analyzed for the 13 short-cert. and 77 long-cert. individuals who had been interviewed in Walter's (1987) study of post-Willie M Program adjustment. Walter used the Diagnostic Interview Schedule, Version III-A (Robins & Helzer, 1985) to generate self-reports of the age at which each individual (a) first got into trouble at school, (b) first had a serious fight, (c) first stole something, (d) first intentionally damaged property, and (e) first was arrested as a juvenile. We compared the groups on the *average* age of first commission of antisocial acts (averaging across the various antisocial categories) and on the *lowest* age at which *any* of the antisocial acts was first committed. As Table 1 shows, short- and long-cert. group means were quite similar on both measures.

Arrest Data

We collected arrest information for all of our subjects from the State Bureau of Investigation (SBI) for the period September 1981 through November 1987. We considered using data on convictions, but such data were less accurate as an index of the timing of criminal activity. Conviction records lagged well behind the incidents that provoked arrest, partly because of delays in trial dates and partly because of frequent plea bargains and delays in pressing charges. So we used arrest records, focusing on one outcome measure: time elapsing between the termination of Willie M services and first arrest. As noted above, we excluded individuals who were unavailable for arrest during some part of the criterion period because they moved out of state, disappeared, or were incarcerated because of a pre-aging-out arrest.

Results

Survival Analysis Procedure

To compare the short- and long-cert. groups' rate and time to arrest, we used survival analysis (Greenhouse, Stangl, & Bromberg, 1989; Miller, 1981). This set of techniques is used to analyze data where length of time to the occurrence of some target event (in the present case, being arrested) is the variable of interest. The period during which subjects are observed for the target event is the *risk period*, and the length of time from the beginning of the risk period to the time at which the event occurs is the *survival time*. Unlike a number of other statistical techniques (e.g., multiple regression), survival analysis can make full use of the data of subjects who never experience the target event, and it allows the use of all subject follow-up data even if the follow-up time periods are not equal for all subjects (see Greenhouse et al., 1989). Survival analysis is also superior to a simple chi-square comparison of the proportion of different groups experiencing the target event because it takes into account the time that transpires before the target event.

In the present study, being arrested subsequent to the end of Willie M services was the target event. The length of time from service termination to the first arrest was the survival time. Although 4- to 5-year follow-up data were available for some subjects (i.e., those who finished the program in the early

Table 1
Characteristics of Short-Cert. and Long-Cert. Groups

Variable	Short-cert.		Long-cert.		Statistical test
	M	*SD*	*M*	*SD*	
Months of services received					
Inpatient child therapy	0.08	0.29	1.44	3.16	$t(153) = 4.93^*$
Outpatient child therapy	0.67	1.15	7.82	8.06	$t(154) = 9.53^*$
Outpatient family therapy	0.58	0.79	5.40	5.82	$t(154) = 8.97^*$
Child day treatment	0.00	0.00	2.09	4.03	$t(154) = 6.23^*$
Supervised group living	0.17	0.39	4.13	6.96	$t(154) = 6.72^*$
Supervised independent living	0.00	0.00	1.23	3.49	$t(143) = 4.23^*$
Supervised vocational placement	0.00	0.00	2.16	4.29	$t(154) = 6.07^*$
Demographic and clinical variables					
Age at certification	17.93	0.20	15.90	0.86	$t(166) = 10.79^*$
Percentage non-White	42.86		48.98		$\chi^2(1, N = 168) = <1$
Percentage female	19.05		25.17		$\chi^2(1, N = 168) = <1$
Number from each mental health administrative district					$\chi^2(3, N = 168) = 1.80$
West	7		31		
North Central	5		47		
South Central	6		50		
East	3		19		
IQ	72.21	18.5	72.34	20.5	$t(162) < 1$
DSM-III percentage psychotic diagnosis	0		10.66		$\chi^2(1, N = 134) = 1.42$
DSM-III percentage conduct disorder diagnosis	41.67		48.78		$\chi^2(1, N = 134) = <1$
Age first antisocial act	12.09	2.78	11.55	2.95	$F(1, 88) < 1$
Earliest age first antisocial act	9.62	4.01	9.75	3.38	$F(1, 88) < 1$

Note. Short-cert. = received services less than 3 months; long-cert. = received services more than 1 year. *DSM-III* = *Diagnostic and Statistical Manual of Mental Disorders*, 3rd. ed.
$^* p < .0001$.

1980s), to maintain sufficient subject density we used a 2-year risk period.

Survival Curves for Short- and Long-Cert. Groups

The cumulative survival function represents the proportion of subjects in each group surviving arrest by the length of time from last services. This function is shown in Figure 1. The proportion of subjects in each group who were arrested in each of 10 equally spaced time periods is presented in Table 2. As the table shows, the percentage of long-cert. individuals arrested remained relatively similar across the various time periods, whereas the percentage of short-cert. individuals arrested was more variable. Although the Figure 1 curve for the short-cert. group is somewhat lower than that for the long-cert. group, indicating a higher arrest rate for the short-cert. group, the apparent difference did not approach statistical significance, Mantel-Cox (1) = 0.73, $p = .39$.

Because the curves suggested that there might be a stronger effect during the early part of the postprogram risk period, we analyzed the data separately for two early phases of the risk period: 100 days and one year. These must certainly be regarded as secondary analyses, because they risk capitalizing on chance patterns in the data. At 100 days, when we directly compared the rates of arrest in the short- versus long-cert. groups, we found no significant overall group difference (Fisher's exact test $ps > 0.14$); a comparison of short- and long-cert. survival functions also revealed no significant difference (Mantel-Cox < 2.4, $p > 0.12$). Tests at one year also revealed no significant overall group difference (FET $p > 0.50$) and no difference in survival functions (Mantel-Cox < 1.0, $p > 0.30$).

The three nonsignificant group differences between survival functions were evidently not due to low power; the power of each comparison to reveal a difference of 25% in arrest rates between the two groups was approximately 0.80 (Freedman, 1982).

By the end of the 2-year risk period of interest in this study, the status of the two groups was as follows: Of the short-cert. group of 21, 33% had been arrested, and 67% had survived for the full 2 years without arrest. Of the 147 in the long-cert. group, 25% had been arrested, 63% had survived for 2 years without arrest, and 12% were "censored" (i.e., they had not been in the sample for 2 years when our follow-up period ended).

Secondary Analysis Using Only "Appropriately Served" Long-Certs.

We carried out a second round of analysis using a procedure designed to maximize the advantage of the long-cert. group in the comparison. In this second survival analysis, we included only those former class members who had been classified by program staff at the time of termination as "receiving appropriate services." To merit this classification, the Willie M office requires that "A class member must be receiving services which

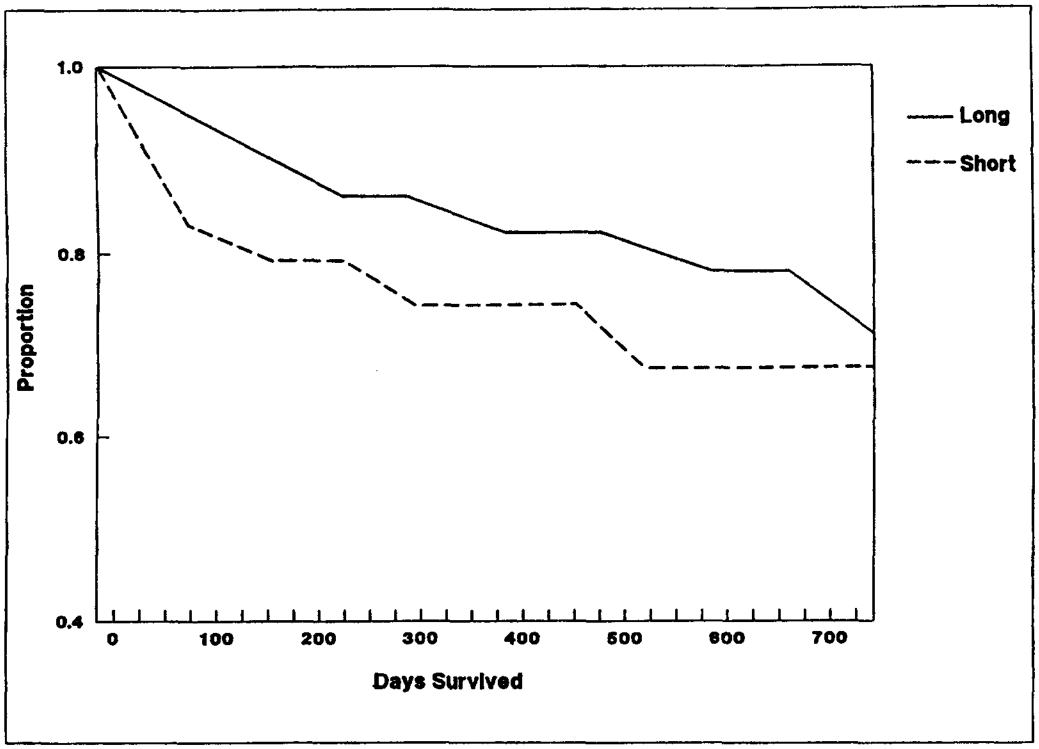

Figure 1. Survival functions for short-cert. and long-cert. groups, full sample included.

adequately address all of his/her treatment and education needs in the least restrictive setting possible."

We made this a secondary analysis because of the following concerns about relying on the appropriateness ratings: (a) The reliability and validity of these ratings are unknown; (b) the requirement that the "appropriate" designation be given only when *all* of the individual's treatment and education needs are met, *and* in the least restrictive setting possible, may mean that the most severe and difficult cases, those who present the most difficult treatment and setting requirements, are excluded from the "appropriate" category, thus biasing results in the direction of positive program effects; and (c) because the appropriateness judgments are made by individuals who have access to information about the class members' *response* to treatment, it is possible that, in some cases, a poor response may increase the likelihood of an "inappropriate" classification, thus excluding unsuccessful cases and further biasing results in the direction of positive program effects.

Despite these concerns, the secondary analysis seemed potentially useful as a test of program effects under conditions

Table 2
Percentage Arrested at End of 10 Equal Time Periods

	Short-cert. group		Long-cert. group	
Days	Number entering time period	% arrested	Number entering time period	% arrested
0–72	21	14.2	147 (102)	3.4 (1.0)
73–145	18	5.5	142 (101)	3.5 (1.0)
146–218	17	0.0	137 (100)	5.8 (5.0)
219–291	17	5.8	129 (95)	0.0 (0.0)
292–364	16	0.0	129 (95)	4.7 (4.2)
365–437	16	0.0	122 (90)	1.6 (2.2)
438–510	16	12.5	120 (88)	1.7 (2.3)
511–583	14	0.0	116 (86)	2.6 (2.3)
584–656	14	0.0	101 (73)	3.0 (2.8)
657–729	14	0.0	96 (69)	3.1 (1.4)

Note. Short cert. = received services less than 3 months; long-cert. = received services more than 1 year. Numbers in parentheses refer to the secondary analysis, in which the long-cert. sample includes only cases classified as appropriately served by program officials.

that most strongly favored the long-cert. over the short-cert. group. To magnify the advantage to the long-cert. group, we excluded those not rated as appropriately served from the long-cert. group, but we retained such cases in the short-cert. group (only 8 of the short-cert. cases were classified as appropriately served). Thus, the secondary analysis included all 21 short-cert. cases, but the reconstituted (R) long-cert. group was reduced from 147 to 102 (i.e., only 69% of the original long-cert. group were classified as appropriately served). The short-cert. group averaged 26.24 days in the program, the R-long-cert. group 922.91.

Comparability of the groups. We compared the short and R-long groups on all the variables shown in Table 1. As in the original analyses, the comparisons showed highly significant differences on all categories of services received and, of course, on age at certification (all *ps* < 001), but no significant differences on race, gender, or regional composition of the groups, or on IQ, *DSM-III* psychosis diagnoses, *DSM-III* conduct disorder diagnoses, or the two "age at first antisocial act" variables (all *ps* > .25).

Survival curves for short- and R-long-cert groups. Figure 2 shows the cumulative survival function for the two groups. The proportion of subjects in each group who were arrested in each of 10 equally spaced time periods is presented in Table 2, with the R-long-cert. data shown in parentheses. Although the Figure 2 curve for the short-cert. group is somewhat higher than that for the long-cert. group, indicating a higher arrest rate for the short group, and the gap between the short and R-long groups is greater than in our original group comparison, the apparent difference was not statistically significant, Mantel-Cox (1) = 2.32, $p = 0.13$. This nonsignificant group difference was evidently not due to low power; as in our original analysis, the power of this comparison to reveal a difference of 25% in arrest rates between the two groups was approximately 0.80 (Freedman, 1982). At the end of the 2-year risk period, 20% of the R-long group had been arrested, 67% had not been arrested, and 14% were censored (see above).

Additional Duration-of-Services Analyses

We undertook two additional analyses to test for possible effects of duration of services. First, we identified the 25% of the long-cert. group who had the longest period in the program (i.e., a long-long group) and compared these with the short-cert. group. Survival functions for the two groups did not differ significantly, Mantel-Cox (1) = 2.11, $p > 0.10$. Second, we compared the long-long group to the 25% of the long-cert. group who had the shortest time in the program (i.e., a short-long group). A comparison of the survival functions for the long-long and short-long groups revealed no significant difference, Mantel-Cox (1) = 0.19, $p > 0.60$. Thus it did not appear that length of time in the program had a significant impact on arrests among the long-cert. individuals.

Male-Only Analysis

Although there was not a significant difference in the male–female ratio of the short- and long-cert. groups, it is certainly possible that there might be sex differences in effects of the program. The small number of Willie M females in the short-cert. group precluded a direct test of such a sex difference; however, an indirect perspective was provided by a comparison of survival functions for male-only short- and long-cert. groups. This comparison, shown in Figure 3, revealed quite similar curves, Mantel-Cox (1) = 0.13, $p > 0.70$. The fact that the curves seem noticeably more similar for male-only samples than for mixed-gender samples suggests a possibility (perhaps testable in future research) that program effects may be more likely to be found among female than male class members.

Group Comparison on Types of Crime

Were different levels of exposure to the program associated with different types of crime after aging out? To address this question, we grouped first arrests during the risk period into two categories studied in previous Willie M research (Weisz, Martin, Walter, & Fernandez, 1990): personal crimes (e.g., assault, armed robbery) and property or victimless crimes (e.g., burglary, drug offenses). There was a higher percentage of personal crime arrests among the short-cert. group (86% of first arrests) than among the long-cert. group (54% of first arrests), but the group difference was not significant, Fisher exact test $p > 0.20$.

Discussion

As noted at the outset, the nature of the Willie M Program, as well as the court consent decree that established it, means that the question of program effectiveness can only be addressed from an oblique angle. Here we addressed the question by comparing (a) individuals who were certified for brief periods and received minimal services with (b) individuals who were certified for longer periods and received more substantial services. The comparison focused on the capacity of the two groups to "survive" without arrest following termination of their services. Although the long-cert. group did show a somewhat more favorable survival curve than the short-cert. group, the group difference did not approach statistical significance; and by the end of the 2nd year, the curves had converged substantially. In a secondary analysis, we tilted the group comparison in favor of the long-cert. group by excluding all those (i.e., 31%) who had not been designated by program officials as appropriately served. In this analysis, the evidence favoring program effects was stronger but still not statistically significant. Thus, the study did not provide strong evidence that the Willie M Program significantly reduces the risk of later arrest among violent and assaultive youth.

Could this null result have been caused by inadequate power to detect group differences? This seems unlikely. Power tests (see above) indicated that the design and sample, in both the primary and secondary analyses, afforded a power of 0.80 to detect a difference of 25%, or 0.75 to detect a difference of 20%, in survival between the short- and long-cert. groups.

Could the null result have been caused by the fact that our follow-up period was limited to 2 years? Certainly it is possible that group differences would have been more pronounced with a more extended follow-up. On the other hand, 29% of the uncensored sample in the primary analysis had been arrested

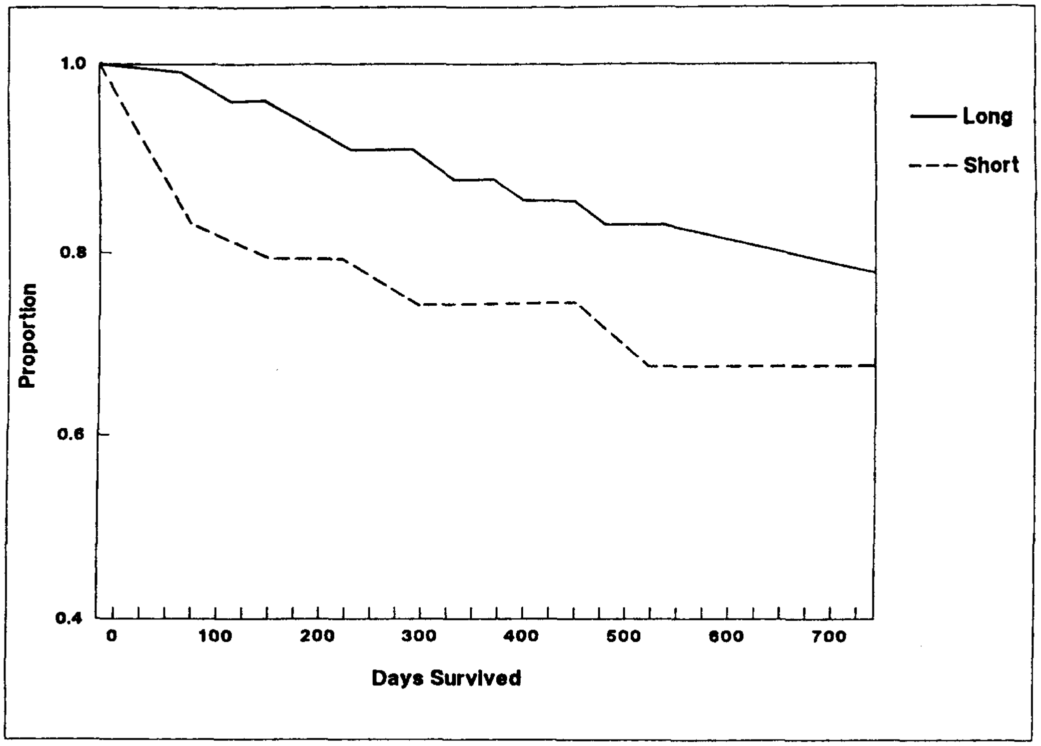

Figure 2. Survival functions for short-cert. group and reconstituted long-cert. group (i.e., with only "appropriately served" individuals included).

at least once by the end of the 2-year follow-up, which suggests that our follow-up period was adequate to detect substantial arrest activity.

It could be argued that arrest data do not provide the most sensitive index of success for the Willie M Program. This argument could take two forms. First, some might maintain that criminal activity would be better assessed by self-reports of the former class members than by official arrest records. Self-reports of delinquent and criminal behavior are viewed by many researchers as more sensitive to change than arrest records. Second, some might argue that criminal activity should not even be the focus of an outcome assessment with this group (i.e., that personality functioning, psychological adjustment, or social relationships might be more sensitive to effects of the mental health aspects of the intervention). On the other hand, reductions in criminal behavior would appear to constitute a particularly important benefit to society, and to former Willie M class members. Moreover, it could be argued that significant positive changes in personality, adjustment, or social relations should be reflected to some extent in such real-life indicators as arrests. Parenthetically, some who are familiar with the program and its origins have argued that *no* outcome assessment is appropriate, inasmuch as the court order that established the program required only the delivery of services, not the production of effects on the class members' behavior. This argument seems problematic, because the delivery of services would be of little value if those services were to produce no effects.

A fourth possible interpretation of the findings relates to the fact, noted earlier, that no *true* control group exists for comparisons with individuals who were treated in the Willie M Program. Any intervention received by our short-cert. group was certainly brief, and modest in scope relative to what the long-cert. group received; however, the efforts of case managers for the short-cert. group to coordinate services over 1 or 2 months of certification may conceivably have had some positive impact. Similarly, the evaluation process and the experience of being identified "Willie M" may have had beneficial effects on both short- and long-cert. subjects. Thus, it is possible that our findings of no group difference result not from an absence of program effects but rather from the fact that both the short- and long-cert. groups profited from the program, and equally so. This would suggest a model of program effects in which benefits are associated with membership in the Willie M class, per se, but are not directly related to the *extent* of services within the program. We do not propose such a model, but it may merit examination. Of course, if such a model were supported by the data, suggesting that 2 months of treatment are as helpful as 2 years or more, this would raise questions about the cost-effectiveness of an extended program.

It must be noted that the Willie M Program was in its early years during the period when all of the subjects in the present study were certified and treated. The program had been in existence just over 4 years when the last of the present subjects aged out. It is possible that, because of the relative youth of the program, the services received by our long-cert. subjects were not appropriate to their needs. Arguing against this possibility are the findings of our second round of analyses. When we used the State's own system for judging whether individuals had

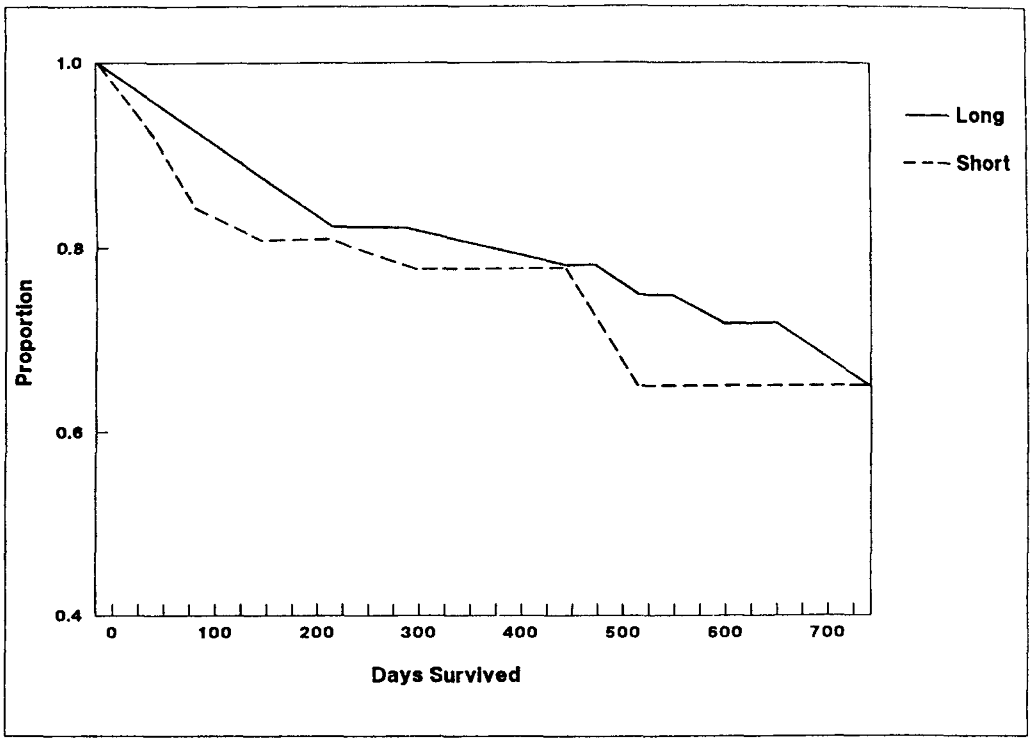

Figure 3. Survival functions for young men in the short-cert. and long-cert. groups.

been appropriately served, and we excluded from the long-cert. sample all those designated as not appropriately served (while retaining in the short-cert. group all those so designated), we still found no reliable difference between the long-cert. and short-cert. survival curves. However, the difference between the curves was more substantial in this latter analysis than in the first, suggesting that services deemed appropriate by program staff and officials may enhance program effects somewhat. This possibility bears exploration in the future. In such exploration, it would be important to ensure that judgments about appropriateness of service are not influenced by the *severity or treatability* of class members' problems or by the *response* of treated youth to their treatment. If severe problems or failure to respond to treatment were often the basis for classification of young people as inappropriately served, then the exclusion of such cases from later outcome analyses would bias comparisons in the direction of spurious treatment effects.

The fact that our data are based on the early years of the Willie M Program suggests another point. Ideally, as a program matures and is refined, its effectiveness should increase. Perhaps this has been the case with the Willie M Program; perhaps program effects are more marked now than in earlier years. Program modifications are limited somewhat by the court consent decree that established program structure and procedures. But it may well be that delivery of services has been improved in a number of ways since the time the subjects in our sample were in the program. One change that has clearly taken place is that Willie M class members are currently being certified and treated at younger ages than in the early years of the program. Some data from longitudinal research (see Olweus, 1979, 1987; Patterson, DeBaryshe, & Ramsey, 1989) and from Willie M client interviews reported by Walter (1987) point to the possibility that earlier intervention may enhance prospects for program effects. Thus, changes over time, both in program characteristics and client age, suggest a need to test for positive effects in later years of the program.

In addition to the alternative interpretations suggested above, one direct interpretation of our findings must be considered: Perhaps the Willie M Program, as structured under the original court consent decree, does not reduce the risk of later arrest to a significant extent. Although the individualized case management provided through the Willie M Program seems reasonable in theory, its effectiveness would certainly depend on whether the case manager has an array of effective services from which to choose. Neither the present findings nor the evidence reviewed in the introduction provide a clear indication that a rich array of such services exists.

Program planners, in this and other programs, might wish to consider experimental additions to existing services. It might prove useful, for example, to develop several precisely focused model treatment programs based on interventions for which success has been documented in treatment outcome research (for relevant reviews and reports, see Casey & Berman, 1985; Garrett, 1985; Kazdin, 1985, 1988; Patterson, Chamberlain, & Reid, 1982; Weisz et al., 1987). Such model programs could be added to currently existing services for randomly selected groups of youth, to assess whether the addition enhances program effects. Over time, such comparisons might enhance our knowledge of what works, and what does not, with such extremely high-risk youngsters.

Clearly, our society needs effective treatment programs for emotionally disturbed violent youth. In developing the Willie M Program, officials in North Carolina have recognized this need and committed significant resources to it. What may now remain is the task of identifying specific interventions that are (a) demonstrably effective with these high-risk youth and (b) sufficiently well-defined that they can be applied appropriately by multiple mental health workers in multiple settings.

References

American Psychiatric Association. (1980). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.). Washington, DC: Author.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (in press). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology.*

Barrett, C. L., Hampe, I. E., & Miller, L. (1978). Research on psychotherapy with children. In S. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (pp. 139–189). New York: Wiley.

Behar, L. (1985). Changing patterns of state responsibility: A case study of North Carolina. *Journal of Clinical Child Psychology, 14,* 188–195.

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin, 98,* 388–400.

Curry, J. F., Pelissier, B., Woodford, D. J., & Lochman, J. E. (1988). Violent or assaultive youth: Dimensional and categorical comparisons with mental health samples. *Journal of the American Academy of Child and Adolescent Psychiatry, 27,* 226–232.

Davidson, W. S., Gottschalk, R., Gensheimer, L., & Mayer, J. (1984, May). *Interventions with juvenile delinquents: A meta-analysis of treatment efficacy.* Paper presented at the Joint NIMH/NIJ Conference on Crime and Substance Abuse, Washington, DC.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine, 1,* 121–129.

Garrett, C. J. (1985). Effects of residential treatment on adjudicated delinquents: A meta-analysis. *Journal of Research in Crime and Delinquency, 22,* 287–308.

Greenberg, D. F. (1977). The correctional effects of corrections: A survey of evaluations. In D. A. Greenberg (Ed.), *Corrections and punishment* (pp. 95–158). Newbury Park, CA: Sage.

Greenhouse, J. B., Stangl, D., & Bromberg, J. (1989). An introduction to survival analysis: Statistical methods for analysis of clinical trial data. *Journal of Consulting and Clinical Psychology, 57,* 536–544.

Kazdin, A. E. (1985). *Treatment of antisocial behavior in children and adolescents.* Homewood, IL: Dorsey Press.

Kazdin, A. E. (1988). *Child psychotherapy: Developing and identifying effective treatments.* New York: Pergamon Press.

Lipsey, M. W. (1988). Juvenile delinquency intervention. In H. S. Bloom, D. S. Cordray, & R. J. Light (Eds.), *Lessons from selected program and policy areas: New directions for program evaluation* (No. 37, pp. 63–84). San Francisco: Jossey-Bass.

Lipsey, M. W. (in press). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In Russell Sage Foundation (Eds.), *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Loeber, R. (1988). Natural histories of conduct problems, delinquency, and associated substance use: Evidence for developmental progressions. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 2, pp. 73–124). New York: Plenum Press.

Loeber, R., & Stouthamer-Loeber, M. (1987). Prediction. In H. C. Quay (Ed.), *Handbook of Juvenile Delinquency* (pp. 325–382). New York: Wiley.

Macbeth, G. (1985). North Carolina's Willie M Program: A current perspective. *Popular Government, 50,* 32–39.

Martinson, R. (1974). What works? Questions and answers about prison reform. *Public Interest, 10,* 22–54.

Miller, R. G. (1981). *Survival analysis.* New York: Wiley.

North Carolina Department of Human Resources. (1989). *Report to the Governor and the General Assembly on the Willie M Program, 1988–89.* Raleigh, NC: North Carolina Department of Human Resources.

Olweus, D. (1979). Stability of aggressive reaction patterns in males: A review. *Psychological Bulletin, 86,* 852–875.

Olweus, D. (1987). Bullies [Television interview]. *20/20,* November 27, 1987.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). A developmental perspective on antisocial behavior. *American Psychologist, 44,* 329–335.

Patterson, G. R., Chamberlain, P., & Reid, J. B. (1982). A comparative evaluation of a parent training program. *Behavior Therapy, 13,* 638–650.

Robins, L. N., & Helzer, J. E. (1985). *Diagnostic Interview Schedule: Version III-A.* St. Louis, MO: Department of Psychiatry, Washington University School of Medicine.

Walter, B. R. (1987). *Assessing the adjustment of aged-out class members from the Willie M Program.* Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

Weisz, J. R., Martin, S., Walter, B. R., & Fernandez, G. A. (in press). Differential prediction of young adult arrests for property and personal crimes: Findings of a cohort follow-up study of violent boys from North Carolina's Willie M Program. *Journal of Child Psychology and Psychiatry.*

Weisz, J. R., Weiss, B., Alicke, M., & Klotz, M. (1987). Effectiveness of psychotherapy with children and adolescents: A meta-analysis for clinicians. *Journal of Consulting and Clinical Psychology, 55,* 542–549.

Whitehead, J. T., & Lab, S. P. (1989). A meta-analysis of juvenile correctional treatment. *Journal of Research in Crime and Delinquency, 26,* 276–295.

Received April 17, 1990
Revision received June 8, 1990
Accepted July 2, 1990 ■