Published in *Anthrozoos*, (1998),11, 194-200.

Dolphin-Assisted Therapy: Flawed Data, Flawed Conclusions

Lori Marino, Ph.D.

Neuroscience and Behavioral Biology Program

Emory University

Atlanta, Georgia 30322

Scott O. Lilienfeld, Ph.D.

Department of Psychology

Emory University

Atlanta, Georgia 30322

Dolphin-Assisted Therapy: Flawed Data, Flawed Conclusions

Two reports on the short-term and long-term effectiveness of dolphin-assisted therapy (DAT) for children with severe disabilities have recently appeared in this journal (Nathanson, Castro, Friend, & McMahon, 1997; Nathanson, 1998). The authors of these reports concluded that the administration of DAT to severely disabled children: (1) "significantly increases motivation, attention span, gross and fine motor skills, and speech and language" (Nathanson et al., 1997, p. 97), (2) achieves positive results more quickly and more cost effectively than conventional long-term therapy (Nathanson et al., 1997), and (3) produces increases in functioning that are maintained or improved upon for at least one year (Nathanson, 1998). Nevertheless, a methodological analysis of these studies demonstrates that these conclusions do not withstand careful scrutiny.

Before discussing methodological issues, it is important to point out that the authors' theoretical rationale for DAT is dubious at best. Specifically, Nathanson and colleagues' contention that an attention deficit "explains why disabled populations have such difficulty with learning and motivation" (Nathanson et al., 1997, p. 91; Nathanson, 1998, p. 23) is inconsistent with our current understanding of almost all of the disorders (e.g., infantile autism, Cri-du-chat syndrome, cerebral palsy, Tuberous sclerosis) afflicting the subjects in their studies. Would Nathanson and his colleagues have us believe that children with Cri-du-chat syndrome, for example, would have essentially normal IQs if only they could learn to focus their attention? Moreover, if the attention

deficit hypothesis were correct, it would follow logically that individuals with attention deficit-hyperactivity disorder (ADHD), for whom attentional problems are a core deficit (Douglas & Peters, 1979), should be severely intellectually disabled. To the contrary, the overall IQ scores of children with ADHD are either not significantly different from normal samples (Anastopoulos, Spisto, & Maher, 1994; Carter, Zelko, Oas, & Waltonen, 1990) or are only slightly below normal (Farone, Biederman, Krifcher Lehman, Spencer, Norman, Seidman, Kraus, Perrin, Chen, & Tsuang, 1993). Furthermore, despite their claim that attention deficits underlie both their subjects' disabilities and the effectiveness of DAT, Nathanson et al. (1997) never assessed attention in their subjects either before or after DAT. Nor did they examine whether improvements in attention were correlated with improvements on their dependent measures. Therefore, there is no way to assess the validity of their theoretical rationale, because it was never put to a test.

In addition to being based on an implausible hypothesis, the studies by Nathanson et al. (1997) and Nathanson (1998) are seriously flawed on methodological grounds. Table 1 displays the principal threats to the validity of psychological studies (see Cook & Campbell, 1979; Kendall & Norton-Ford, 1982; Shaughnessy & Zechmeister, 1994) that undermine the credibility of Nathanson et al. (1997) and Nathanson (1998).

TABLE 1 ABOUT HERE

As Cook and Campbell (1979) noted, the presence of even one major threat to validity renders a study's findings questionable or even uninterpretable. As Table 1 shows, Nathanson et al. (1997) and Nathanson (1998) violated several important criteria for

validity. Most of these threats relate to internal validity, i.e., the methodological soundness of the study, but some are also relevant to external validity, i.e., generalizability of the findings. Because of space constraints, we limit ourselves only to the most serious threats to validity present in both studies. Before discussing these methodological flaws, a brief overview of Nathanson and colleagues' first study is necessary.

Nathanson et al. (1997) compared the effectiveness of a 2-week DAT program with a 6-month conventional physical and speech therapy regime in children with multiple disabilities of varying etiologies. Each participant had received at least 6 months of conventional therapy immediately prior to DAT and was assigned (nonrandomly) to either a physical treatment-goal group or a speech treatment-goal group, depending upon their disability and availability. This study utilized an approximation of a pre-post design, and the participation of all children was contingent upon their inability to respond independently on either a physical or verbal task prior to DAT. During 2 weeks (17 sessions) of DAT, all children were assessed for their ability to respond independently on the same task. This assessment constituted the "post-test" score and all improvements were attributed to the effects of DAT. The results indicated that 57% - 71% of the children (depending on the group) were able to make the independent response during DAT compared with 0% prior to DAT. The authors used this finding to argue that DAT is an effective treatment for severe disabilities and is markedly more beneficial and efficient than conventional therapy. These conclusions are unwarranted for the following reasons.

Although Nathanson et al. (1997) discussed in detail the advantages of single-subject designs in therapy outcome research (pp. 92-94) and asserted that "a series of single subject, multiple baseline across settings experimental design was used" (p. 93), these comments are misleading. In fact, Nathanson et al. (1997) never presented single-subject data or discussed findings at a single-subject level. All data were presented and analyzed in aggregate form and therefore do not permit examination of change within individuals. This omission is highly problematic, because improvement at an overall level may mask substantial heterogeneity in subject outcomes. Indeed, it is conceivable that a subset of children in Nathanson et al. (1997) became worse following DAT, but that their deterioration was offset by those children who improved. Regrettably, Nathanson et al.'s analyses do not permit the reader to evaluate this possibility.

The source of most of the major flaws in Nathanson et al. (1997) is the absence of experimental control, making it impossible to determine whether their results were due to the specific effects of DAT or to a host of potentially confounding factors, such as the experience of being in water. Nathanson et al. (1997) did not utilize a no-treatment control group or a control group of individuals exposed to an alternative intervention. Nor did they employ a dismantling strategy (Kazdin & Wilson, 1978) to systematically expose subjects to different treatment components (e.g., interaction with dolphins, interaction with trainers, immersion in water). Neither did they use pre-test/post-test counterbalancing techniques (Shaughnessy & Zechmeister, 1994) to examine the potential influence of order effects. Nathanson et al. (1997) dismissed such problems by appealing to the study by Nathanson and de Faria (1993), who compared the effects of in-water learning *with* dolphins and without dolphins (with children's' favorite toys used *in*

lieu of dolphins) in order to assess the relative effects of in-water therapy alone. Nathanson and de Faria (1993) reported greater effectiveness of in-water therapy with dolphins compared with in-water therapy without dolphins. Nevertheless, their study is seriously flawed. When comparing subjects' responses with dolphins versus favorite toys, the two conditions took place at entirely different facilities, viz., The Dolphin Research Center versus a local motel, resulting in a complete confounding of treatment condition with setting. Therefore, despite Nathanson et al.'s (1977) claims that the results of Nathanson and de Faria (1993) negate the need for control groups in subsequent studies of DAT, this conclusion is unjustified. Without a control group in Nathanson et al. (1977), there is no way to determine whether subjects' post-DAT responses were due to the specific effects of DAT, to a placebo effect (see Table 1), or to such nonspecific factors as increased interpersonal contact, increased interpersonal attention, or a number of other plausible variables.

One particularly troublesome confound in Nathanson et al. (1997) is novelty. The authors claimed to control for novelty by discounting any of the subjects' responses as independent, i.e., meeting the treatment criterion, until after the fourth session. Rather than controlling for novelty, this procedure renders claims concerning the effectiveness of DAT all the more difficult to evaluate. Specifically, it is not possible to determine whether any subjects responded independently very early in the treatment phase – a result that would suggest the possibility of a novelty effect.

A further threat to validity resulting from the absence of a control group is history (Cook & Campbell, 1979), i.e., the occurrence of events outside of therapy that can have an effect on the dependent measures. For example, Nathanson et al. (p. 91) noted that

many of the children treated with DAT came from different states and even different countries. Most or all of these children surely encountered a plethora of new experiences during the course of DAT: travel to a novel and exciting environment, an extended stay at a hotel or unfamiliar lodging, meeting new people, interacting with other children, and so on. Although some of these experiences may have remained relatively constant over the course of treatment, others almost certainly did not. Without a randomized control group, it is impossible to ascertain whether any of these extra-therapy events might have contributed to improvement on dependent measures.

Moreover, Nathanson et al.'s (1997) design, which provided children with repeated practice on both verbal and motor stimulus materials and then tested them on the same stimuli on which they had practiced, is subject to the validity threat of testing (Cook & Campbell, 1979). Because Nathanson et al. did not examine whether subjects' knowledge and skills generalized to words or motor tasks on which they had not been explicitly tested, practice effects cannot be ruled out as an explanation for their primary findings. Although we do not wish to imply that improvements on these tasks are of minimal clinical significance, it is crucial to note that Nathanson et al.'s claims concerning the effectiveness of DAT are not limited to the specific stimuli used in their study, but instead extend to language and motor skills in general (e.g., see p. 97).

Another major set of flaws inherent to Nathanson et al. (1997) concerns how the subjects' responses were measured and elicited. Both of these flaws introduce the possibility of experimenter expectancy effects. Because the interns who scored the subjects' behavior were aware of the desired outcome, the objectivity of the scoring procedure is suspect. Nathanson et al. might argue that the criterion responses, e.g.,

placing a ring on a peg, were so clear-cut that no bias in scoring was possible. Nevertheless, the fact remains that the criterion response involved a categorical distinction between assisted and independent responses. Because there were apparently no rigorous criteria for distinguishing assisted from independent responses, subtle interpretative bias may have occurred. In addition, the possibility of subtle and unintentional cueing of subjects by the therapists is difficult to exclude. A large body of research shows that experimenter expectancies can influence not only how subjects' responses are coded and interpreted, but even the responses themselves (Rosenthal, 1994).

These concerns are exacerbated by the fact that Nathanson et al. (1997) are unclear in reporting if and how they measured inter-rater reliability. On the basis of high inter-rater reliabilities from a previous study (Nathanson & de Faria, 1993), they stated that "For purposes of data analysis in the current investigation, inter-rater reliability was 1.00 on measured independent responses" (p. 95). It is not clear from this statement whether Nathanson et al. (a) based this inter-rater reliability measure on all trials in the present study, (b) chose to assume a reliability of 1.0 based on the previous study, or (c) only included trials on which there was perfect inter-rater agreement. Without such information, it is impossible to gauge the reliability, and therefore validity, of Nathanson et al.'s dependent measures. Moreover, even if high inter-rater reliabilities were obtained by Nathanson et al. (1997), their raters could not have been blind to condition because there was no control group. As a consequence, these raters' errors may have been systematic rather than unsystematic.

The absence of a control group also renders regression (Cook & Campbell, 1979) an especially serious threat to validity. Regression, which refers to the tendency of extreme scores to become less extreme upon retesting, is of particular concern in pre-post designs (Kendall & Norton-Ford, 1982). In addition, the problem of regression is typically compounded in treatment outcome studies, because individuals often bring themselves (or are brought) to treatment when their condition is at its worst. The failure to take regression into account may lead the investigator to fall prey to the regression fallacy, which is the error of attributing improvement to the intervention, rather to statistical regression (Gilovich, 1991).

The interpretation of Nathanson et. al.'s findings is further complicated by the confound of instrumentation (Cook & Campbell, 1979), i.e., a change in the assessment of the dependent variable at different points in the study. Although Nathanson et al. gave all subjects a pretest score of 1.0 (capable of a response only with assistance) on the basis of written reports, parent interviews, and direct observations, they assessed the "post-test" score in an entirely different way. Specifically, subjects' responses were counted as successful if they achieved the criterion physical and verbal behaviors at *any point* between sessions 5 and 17. For example, if a child achieved criterion in session 5 but failed to meet this criterion in all subsequent sessions, the outcome would still be counted as successful. Because Nathanson et al. did not provide information regarding the number of children who failed to maintain the criterion level of responding following an initial success, their primary dependent measure is extremely difficult to interpret and does not provide a stringent test of DAT's effectiveness.

It is clear that Nathanson et al.'s (1997) study is seriously deficient from a methodological standpoint. Regrettably, the follow-up to this study, Nathanson (1998), is plagued by a number of the same validity threats (i.e., history, placebo/nonspecific effects, instrumentation, and regression) found in Nathanson et al. (1997) and suffers from many additional validity threats as well. Nathanson (1998) attempted to assess the long-term effectiveness of DAT by sending a survey to parents of children who had participated in either a 1 or 2 week DAT program at least 1 year earlier. On the basis of parents' responses, Nathanson (1998) concluded that: (1) "children maintained or improved skills acquired in therapy about 50% of the time even after 12 months away from therapy" (p. 22), (2) 2 weeks of DAT produced significantly better long-term results than 1 week of DAT, and (3) there were no differences in the long-term effects of DAT as a function of the etiologies of the participants' disorders (genetic, brain damage, unknown cause). These conclusions, like those of Nathanson et al. (1997), are unwarranted.

Because Nathanson's (1998) study lacked a control group, the possibility of history and multiple intervention interference cannot be excluded. These validity threats are especially problematic in studies, like Nathanson (1998), that are long in duration. Nathanson (1998) acknowledged that most of his subjects received conventional therapies following DAT and prior to the parental reports on which his post-test measurements were based. Yet he neglected the fact that it is inappropriate to attribute improvement solely to DAT when subjects received months of conventional therapy between the pre- and post-test measurements. Nathanson (1998) claimed that his questionnaire was valid because it asked parents to assess the "specific behavioral

improvement and maintenance of the behavior as a direct result of dolphin-assisted therapy…" (p. 24). It is not reasonable to assume, however, that parents were able to distinguish between those aspects of their child's behavior that were affected by DAT and those that were influenced by other factors, not the least of which were other treatments. Nathanson also included items on the parents' survey concerning the children's' responses to various forms of conventional therapy (e.g., speech therapy, special education classes). He interpreted improvements in these areas as indicating that DAT "has been able to increase, by more than 50%, the amount of time that children participate in and benefit from conventional therapies" (p.28). Yet he interpreted reports of no improvement on 15% of the behaviors as due to "little or no follow-up in conventional therapy…" (p. 28), among other factors. It appears that when conventional therapy was associated with improvement in functioning, Nathanson attributed this finding to the potentiating effect of DAT on conventional treatments. But when there was no improvement following DAT, he attributed these results to a lack of conventional follow-up therapy. It would be equally plausible to argue that the 85% of behaviors that were maintained or improved following DAT were in fact due to the effects of conventional therapies and other interfering factors.

One of the most problematic threats to validity in uncontrolled long-term outcome studies is subject maturation. Nathanson (1998) claimed that the problems of history and maturation were mitigated by the use of "a large, randomized, heterogeneous (i.e. etiologies) subject pool, and a valid and closed form ratio response scale, with clearly defined behaviors…" (p.29). This argument is a non sequitur. In fact, a more reliable

and valid response scale would only increase the probability of detecting history and maturation, which are genuine, albeit unwanted, effects.

In addition to a lack of control over the validity threats (e.g., maturation and history) intrinsic to an uncontrolled long-term outcome study, Nathanson's (1998) method of assessing DAT's effectiveness renders his results virtually uninterpretable. One of the most dangerous threats to validity is the presence of demand characteristics, i.e. the tendency of participants to alter their responses in accord with what they believe to be the researchers' hypotheses. Not only did Nathanson fail to guard against this problem, but he exacerbated it in two ways. First, each behavioral item on the parental survey was preceded by the statement "As a result of Dolphin Human Therapy, my child has maintained or improved in his/her ability to…" (p.24). Parents were asked to circle one of six responses ranging from "never" to "always" or "does not apply". In an additional open-ended section, parents were invited to list additional behaviors that were maintained or improved as result of DAT. Therefore, the hypothesis of the researcher, namely that DAT is effective, was made virtually explicit to respondents. Second, despite Nathanson's acknowledgment that a valid survey uses items that "account for all possible responses" (p.24), the questions in his survey were limited to inquiries about positive effects of DAT, namely behaviors that were maintained or improved. Remarkably, behaviors that might have worsened were never systematically assessed or analyzed. Moreover, even though "parents were invited to write in general comments about the long-term effects of the program" (p. 25), these comments were not coded or used in the analyses. As a consequence, Nathanson did not follow his own acknowledged prescription for questionnaire validity.

Nathanson's (1998) study is also plagued by the validity threat that Cook and Campbell (1979) referred to as subject mortality. Of 137 questionnaires sent out to parents, only 52% were returned. This relatively low rate of return raises the possibility that parents who responded were unrepresentative of the entire sample of parents whose children were given DAT. Because Nathanson made no attempt to determine if responders differed from non-responders on potentially important variables (e.g., demography, etiology of disability) and, more to the point, short-term DAT outcome, this possibility cannot be evaluated.

Additionally, because all of the outcome data in this study derived from parents, who cannot be assumed to be objective reporters, the results and conclusions of this study are further undermined by potential informant bias. It is well documented that memory is far more reconstructive than has traditionally been thought (Loftus, 1993), and that retrospective reports are often of suspect validity. For example, parents may selectively recall their memories of their children's' improvement in accord with their hopes and expectations. In an elegant series of studies, Ross (1989) showed that individuals in treatment studies often unintentionally distort their memories of improvement on the basis of their expectations concerning change. For example, if individuals expect to improve as a result of treatment but experience no objective improvement, they will often recall their pre-treatment status to be worse than it actually was (Conway & Ross, 1989). The same phenomenon might account for Nathanson's (1998) results: parents who expect improvement following DAT might remember their children's pre-DAT behaviors as worse than they were objectively. In addition, there are a variety of reasons why the parents in Nathanson's (1998) study might have unintentionally distorted their estimates

of their childrens' current functioning. Among these reasons is effort justification, which is the tendency of individuals who expend a great deal of energy, time, and money in a treatment to justify this effort by convincing themselves that this treatment must have been effective (Cooper, 1980).

Finally, for reasons that are unclear, Nathanson (1998) analyzed the data in his Table 1 (p. 26) and his Table 2 (p. 27) in two different ways, despite the fact that the dependent measures in both tables were identical. For Table 1, which presents the mean levels of 15 behaviors rated by parents as maintained or improved, he aggregated these behaviors into a single scale and reported the mean overall level of maintenance/ improvement. Yet for Table 2, which presents the mean levels of the same 15 behaviors as a function of etiology, he did not aggregate these behaviors into a single scale. Instead, he examined the item (item 10) that exhibited the largest difference across groups, conducted an analysis of variance (ANOVA) on this item, and concluded that there were no differences in outcome across the three groups because the ANOVA was not statistically significant. This method of analysis is inappropriate, because it is well known that individual items are extremely unreliable. Nathanson should have either pooled the items into a single scale, as he did for the items in Table 1, or conducted a multiple analysis of variance (MANOVA) across all 15 items. As a consequence, Nathanson's conclusion that DAT is equally effective across etiologies is unjustified. In fact, inspection of Table 2 reveals that the children in Group 1 (genetic abnormalities) showed lower levels of maintenance or improvement than the other two groups on 12 out of 15 items.

In summary, a plethora of serious threats to validity and flawed data analytic procedures render the findings of Nathanson and colleagues uninterpretable and their conclusions unwarranted and premature. Given that Nathanson and de Faria (1993), Nathanson et al. (1997), and Nathanson (1998) are the only peer-reviewed published studies on the effects of DAT, the current evidence for the efficacy of DAT can at best be described as thoroughly unconvincing. Both practitioners of DAT and parents who are considering DAT for their children should be made aware that this treatment has yet to be subjected to an adequate empirical test, and that Nathanson and colleagues' attention deficit hypothesis remains an explanation in search of a phenomenon.

References

Anastopoulos, A.D., Spisto, M.A., and Maher, M.C. 1994. The WISC-III Freedom from Distractibility Factor: Its utility in identifying children with attention deficit hyperactivity disorder. *Psychological Assessment*, 6(4):368-371.

Carter, B.D., Zelko, F.A.J., Oas, P.T., and Waltonen, S. 1990. A comparison of ADD/H children and clinical controls on the Kaufman assessment battery for children (K-ABC). *Journal of Psychoeducational Assessment*, 8: 155-164.

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. Journal of Personality and Social Psychology, 47, 738-748.

Cook, T.D. and Campbell, D.T. 1979. *Quasi-experimentation*: *Design and analysis issues for field settings.* Boston, Mass.: Houghton Mifflin.

Cooper, J. (1980). Reducing fears and increasing assertiveness: The role of dissonance reduction. Journal of Experimental Social Psychology, 16, 199-213.

Douglas, V.I. & Peters, K.G. 1979. Toward a clearer definition of the attentional deficit of hyperactive children. In *Development of Cognitive Skills*, 173-247, eds. G.A. Hale and M.. Lewis, New York: Plenum Press.

Faraone, S.V., Biederman, J., Krifcher Lehman, B., Spencer, T., Norman, D., Seidman, L.J., Kraus, I., Perrin, J., Chen, W.J., and Tsuang, M.T. 1993. Intellectual performance and school failure in children with attention deficit hyperactivity disorder and in their siblings. *Journal of Abnormal Psychology*, 102(4):616-623.

Gilovich, T. 1991. *How we know what isn't so: The fallibility of human reason in everyday life.* New York: The Free Press.

Kazdin, A.E. and Wilson, G.T. 1978.*Evaluation of behavior therapy: Issues, evidence, and research strategies.* Cambridge, Mass.: Ballinger.

Kendall, P.C. and Norton-Ford, J.D. 1982.Therapy outcome research methods. In *Handbook of Research Methods in Clinical Psychology*, 429-460, eds. P.C. Kendall and J.N. Butcher, New York: John Wiley & Sons.

Loftus, E.F. (1993). The reality of repressed memories. American Psychologist, 49, 518-537.

Nathanson, D.E. 1998. Long-term effectiveness of dolphin-assisted therapy for children with severe disabilities. *Anthrozoos*, 11(1):22-32.

Nathanson, D.E., de Castro, D., Friend, H., and McMahon, M. 1997. Effectiveness of short-term dolphin-assisted therapy for children with severe disabilities. *Anthrozoos*, 10(2/3):90-100.

Nathanson, D.E. and de Faria, S. 1993. Cognitive improvement of children in water with and without dolphins. *Anthrozoos*, 6(1): 17-29.

Rosenthal, R. 1994. Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science,* 3: 176-179.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. Psychological Review, 96, 341-357.

Shaughnessy, J.J. and Zechmeister, E.B. 1994. *Research methods in psychology*. New York: McGraw-Hill.

Table 1.

Principal threats to validity in Nathanson et al. (1997) and Nathanson (1998)

Validity threat	Definition	Nathanson et al.(1997)	Nathanson (1998)
Placebo/nonspecific effects	Improvement resulting from the expectation of improvement (placebo effect) or from effects (e.g., increased attention from therapists, increased interpersonal contact) that are common to many or most psychological treatments	X	X
History	The occurrence of potentially therapeutic events other than the intended treatment during the course of the study	X	X
Testing	Improvements in participants' test performance resulting from the effects of testing itself (e.g., practice effects)	X	
Experimenter expectancy effects	The tendency for researchers to unintentionally bias the results of the study in accord with their hypotheses	X	
Regression	The tendency of participants with extreme scores at one time point to obtain less extreme scores upon retesting	X	X
Instrumentation	Changes in the assessment of the dependent measure at different points in the study (e.g., pre-test vs. post-test)	X	X

Multiple intervention interference	The administration of treatments other than the intended treatment during the course of the study	X
Maturation	Changes in participants over time due to naturally-occurring developmental effects	X
Demand characteristics	The tendency of participants to alter their responses in accord with their suspicions concerning the researchers' hypotheses	X
Subject mortality	Unrepresentative loss or drop-out of participants from the original sample	X
Informant bias	The tendency of informants to selectively recall the amount of improvement in accord with their hopes and expectations (retrospective bias), or to unintentionally distort their estimates of improvement as a consequence of effort justification or other factors	X

Note: List of threats to validity partly adapted from Cook and Campbell (1979), Kendall and Norton-Ford (1982), and Shaughnessy and Zechmeister (1994)