DOCUMENT RESUME

ED 286 335 EC 200 515

AUTHOR Shaver, James P.; And Others
TITLE The Methodology and Outcomes of Research on Modifying Attitudes Toward Persons with Disabilities: A Comprehensive, Systematic Review.
PUB DATE Apr 87
NOTE 57p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). Table 1 contains marginally legible print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Attitude Change; *Data Analysis; *Disabilities; Effect Size; *Meta Analysis; Outcomes of Treatment; *Research Methodology

ABSTRACT

Meta-analysis was used to conduct an integrative review of the research on modifying attitudes toward persons with disabilities. Prior reviews, using small samples and lacking systematic data collection and analysis, were unable to draw firm conclusions about the effectiveness of attitude modification techniques. In the hope of correcting these deficiencies, an exhaustive search of all English-language research reports in the United States and Canada was undertaken that identified 667 potentially relevant items. Screening for relevance and adequacy of information reduced the accessible population for the integrative review to 273 reports describing a total of 644 treatment groups. A coding instrument containing some 162 categories was used . Data analysis is detailed for the following: inter-rater reliability, global quality indicators, specific validity threats, measurement concerns, study populations and samples, replications, qualification of results, treatment outcomes, comparison of experimental treatments, treatment variability, and concomitant variables. The major conclusion was that clear-cut indications of the overall efficiency of techniques for modifying attitudes toward disabled persons or of reliable differences in efficacy were not found. (DB)

THE METHODOLOGY AND OUTCOMES OF RESEARCH ON MODIFYING ATTITUDES TOWARD PERSONS WITH DISABILITIES: A COMPREHENSIVE, SYSTEMATIC REVIEW*

James P. Shaver
Utah State University

Charles K. Curtis
University of British Columbia

Joseph Jesunathadas
Utah State University

Carol J. Strong
Utah State University

Legislation and judicial decisions are bringing handicapped persons into the mainstream of educational, social, and economic life in this society. Nevertheless, negative attitudes toward persons with disabilities continue to be detrimental to their potential to live dignified, productive lives and to contribute to society. A major research interest has been how to modify the negative attitudes and thereby mitigate the effects on persons with disabilities. That research literature has been reviewed in the past, but this paper is based on the most comprehensive review to-date (Shaver, Curtis, Jesunathadas, & Strong, 1987).

Prior Reviews of Research

Seven full and eight brief prior reviews of primary research on the modification of attitudes toward disabled persons were located. These reviews were examined for methodological soundness and for their contributions to knowledge using questions developed from the work of Jackson (1978, 1980) and others, with the primary research process as a model.

Although building on prior works is a standard approach for advancing knowledge in a field, most of the reviewers ignored previous, but relevant,

---

Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April 24, 1987.

*This paper is based on a research project funded by the Research in Education of the Handicapped Program, Office of Special Education and Rehabilitative Services, U. S. Department of Education.

reviews. They did not draw on the findings of earlier reviewers; nor did they use inadequacies in prior reviews as a basis for improving the quality of their work. (This discussion is based on Shaver et al., 1987, Ch. 2.)

The methods used to locate and select primary studies were seldom reported, and the possibility of sampling bias was present in each review. The importance of sample selection can be illustrated by comparing the number of primary research reports cited in prior reviews with those identified for our review of literature. The total number of individual attitude change studies cited in the seven reviews and eight brief reviews was 192. The median number of primary studies referenced in the full reviews was 31 ($\overline{X}$ = 38; range, 24-70); in the brief reviews, the median was 11 ($\overline{X}$ = 13; range, 5-27). Our literature search yielded 273 primary research studies that met specific criteria for inclusion in our review of research on the modification of attitudes toward the disabled. An additional 394 studies were discarded as not suitable* for the present review. The limited reference lists and the small number of primary studies that were cited in more than one of the prior reviews cast serious doubt on the representativeness of the samples.

Many of the primary studies cited in prior reviews were low in treatment and internal validity; although this was mentioned in several reviews, it could not be determined how or if such studies were weighted when decisions concerning the effectiveness of particular interventions were reached. It seems apparent, given that lack of discussion, that treatment and internal validity were not explicitly considered in most reviews. Poorly designed and

---

*Of these, 363 were deemed irrelevant because they were, for example, correlational studies, used instruments that did not fit our definition of attitude, or attitudes toward mainstreaming rather than toward disabled persons were assessed. An additional 31 studies were discarded due to lack of information.

executed studies were included in the reviews without examining the association between design quality and outcomes.

Other methodological weaknesses were found in most of the reviews. Primary studies were placed into loosely defined intervention categories, with the result that important differences in sample and intervention characteristics were frequently disregarded. Narrative reports of programs and reviews of literature were cited as though they were primary studies. In several reviews, primary studies were misinterpreted and irrelevant studies were cited. Furthermore, there was a general tendency to report the findings of complex primary studies in simple treatment-outcome terms and, in some cases, to report only partial results. Moreover, even the statistical significance of findings was not presented in most reviews, and none reported research results in an effect size metric independent of sample size. And, studies which failed either to identify the dependent variable or to provide reliability or validity data for project developed instruments appeared to be accepted uncritically. (The prior reviews are discussed in detail in Shaver et al., 1987.)

The most common conclusion in the reviews, as summarized in Table 1, was that there was not adequate evidence to support the effectiveness of any particular approach to attitude change. Although information plus direct contact with disabled persons seemed most likely to have positive effects, even those results were deemed equivocal because of limited samples, poor study designs, and inconsistent results. It was not clear whether the generally indefinite conclusions about the effectiveness of interventions for modifying attitudes accurately reflected the state of available research knowledge or were the result of the limited numbers of studies reviewed and the lack of a systematic approach to data collection and analysis. A

Summary of Reviewers' Conclusions

Author	Contact	Information	Contact + Information	Vicarious/Simulation	Other Comments
Anthony (1972)	Studies with wide variety of disabled persons, no consistent changes (p. 119) Contact in and of itself does not change attitudes significantly (p. 120) May even reinforce neg. attitudes (p. 123)	Regardless of how info. presented, positive effect (p. 120, 121)	Consistently favorable impact (p. 121, 123) Limited research, with college volunteers or trainees in helping professions; lack of data on other age groups, nonvolunteers, and nonhelping professions (p. 123)		Dearth of exp'l studies (p. 120) Need to include behavioral measures (p. 124) Little known about the time needed (varied in length from 9 hrs to 2 yrs. p. 123)
Yaddis (1974)	No substantial results with contact alone (p. 33)	Most studies produced no significant results (p. 32)	Info. and contact tend to produce more significant results (p. 35) But studies poorly designed (p. 35) Most significant studies required extensive contact—often 40 hrs/wk (p. 35)		Cites Anthony (1972) that most Ss were volunteers and college age (p. 36) Most studies lacked good exp'l designs (p. 36)
Donaldson (1980)	Contact per se not effective (p. 505) Structured contact, pos. change (p. 505); unstructured social or prof'l contact, results equivocal (p. 505) Factors in pos. change: (1)-status (age; social, ed-'l, vocational status; helping relation) (p. 505); (2)-disabled don't act in stereotyped manner (p. 507) Short, structured non-stereo. experiences, short term impact (p. 511)	No causal relationship between limited info. and attitude change (p. 509) If info. confirms negative stereotypes, negative effect (p. 511) Studies of courses not helpful because content unspecified and confounded with contact, media exposure, instructor characteristics (p. 508)		Simulation: only 2 studies. Can be effective if can observe reactions of nondisabled persons (p. 508)	Paucity of research: "literature contains relatively few studies" (p. 505) Failure to test theories (p. 529) Behavioral outcomes & long term effects need investigating (p. 512)
Sandler & Robinson (1981)	Effects of contact equivocal (p. 98)	Effects assessed by few researchers (p. 99) Controlled studies needed (p. 100)	Cited 1 study that info and contact together beneficial (p. 101)		
Westwood et al. (1981)	"Results tend to be inconclusive at best" (p. 221)	"Educational programs" produced equivocal results; results "inconclusive at best". Program content unknown (p. 221)		Simulation: "results . . .are inconsistent" (p. 222)	Contact: "earlier" studies didn't produce change (p. 221) Need to study various media (p. 221)
Towner (1984)	Various approaches with different populations equally effective; similar techniques with diff rent disability groups yielded discouraging and contradictory findings. Positive and negative findings, in addition to non-significant results, from contact and information. Modes of presentation, including simulation, didn't produce significant differences (pp. 249-51)				Instrumentation seemed to have no effect (p. 251). Generally paper & pencil techniques (p. 224) Few attempts to address the complexity of attitudes (p. 224) Most reported no theoretical base (p. 224) Findings contaminated by methodological faults (p. 251)
Horne (1985)	Results inconclusive (pp. 156; 163-4) Limited # of studies, differences in treatments, methodology, & disabilities (p. 156-9)	Inconsistent results (pp. 163-4)	More successful, with prof'ls, but still not consistent (p. 151)	Role playing, children's books—results mixed (p. 178)	Interactions rarely assessed (p. 152, 156) Do immediate posttest results hold up? (p. 185)
Fulton (1976)	Contact a factor but not with all social settings	Results with info. equivocal (p. 16)		Role play has potential (pp. 86-7)	Very few experiments that have positively changed attitudes toward physically stigmatized (p. 35)
Johannsen (1979)	Equivocal results (p. 224)	Not much is known about relative effectiveness of techniques (p. 224)			
Rabkin (1972)		Results conflict (p. 167)	Contact with patients and formal instruction effective (p. 166)		Questionnaires, few efforts to measure changes in behavior (p. 163)
Harth (1973)	Social contact not enough (p. 151)	More direct the procedure, the better the results (p. 150)	Effectiveness of knowledge through direct contact supported (p. 160)		No consistent line of research; no theoretical base (pp. 151-2)
Alexander & Strain (1978)*					
Segal (1978)	Can reinforce neg. attitudes if bizarre behavior (p. 215)		"Educated contact" necessary (pp. 215, 216)		
Horne (1979)			Need info. and contact (p. 63)		
Chubon (1982)	Some indication that prof'l experience negatively related (p. 28)				Lack of definition of terms (p. 27) Methodology poor—lack of theory, standardized definitions, refined measurement devices Need to build on findings and experiences of other researchers (p. 27)

*No conclusions based on the research in regard to methods for modifying attitudes toward disabled persons could be found in the article.

comprehensive, systematic, meta-analytic type of study was undertaken to determine which was the case.

Procedures for this Review

Bangert-Drowns (1986) has noted that the choice of a quantitative approach for conducting an integrative review should be based on the purpose for the review. Our intent was to determine what the available research has to say about the effectiveness of treatments or interventions to modify attitudes toward persons with disabilities. For that reason, we adopted the approach to integrating the results of prior research that has been labeled by Glass (1976, 1977) as "meta-analysis". Properly implemented, the meta-analysis approach meets all of the criteria for high quality integrative reviews proposed by Jackson (1980). In conducting a meta-analysis, the reviewer: (1) locates either all studies or a representative sample of all studies on the defined topic; (2) converts the findings of each study, regardless of study quality, to a common metric--that is, computes an effect size for each relevant finding; (3) codes the various characteristics of each study that might have affected the results (such as type of treatment, methodological quality, sample attributes, and type of dependent measure); (4) uses statistics to summarize study outcomes (effect sizes) and to examine the covariations of outcomes and study characteristics; and, (5) draws conclusions based on the results of those analyses.

The Accessible Population of Studies

The purpose of this study was to conduct a comprehensive integrative review of the literature. The target population was all English-language reports of research identifiable through an exhaustive search conducted in

this country and Canada. There was no sampling procedure and only a few of the identified reports could not be obtained, although some that were relevant had to be discarded because adequate information was not reported. Therefore, the set of primary research reports that was reviewed was an accessible population, not a sample.

Of specific interest were empirical investigations of the effects of interventions, or treatments, on the attitudes of nondisabled persons toward persons with disabilities. Correlational research was excluded. In addition to studies with experimental and quasi-experimental designs, single-group studies that involved a planned intervention and the collection of pretest and posttest data were included. Any research directed toward changing attitudes toward persons with disabilities or handicaps was of interest.

"Disabled or handicapped persons" was defined in terms of conventional special education categories, as reflected in Public Law 94-142, to include: mentally retarded, hard of hearing, deaf, speech impaired, visually handicapped, seriously emotionally disturbed (or, mentally ill), orthopedically impaired, deaf-blind, multi-handicapped, and learning disabled, as well as general categories such as "the disabled", "the handicapped", or "physically disabled". Studies of subjects from populations such as "disadvantaged students", "disruptive students", or "slow learners" were not included.

Attitudes toward disabled or handicapped persons was the dependent variable of interest in identifying and selecting primary reports. It was recognized that, consistent with common definitions (e.g., Triandas, Adamopoulas, & Brinberg, 1984), researchers might consider attitudes (which we defined, to provide context, as "interrelated beliefs about and feelings

toward an object which predispose the person to act in certain ways") as having cognitive, affective, and/or behavioral components. It was also recognized that "attitudes" might be assessed in a variety of ways, including paper-and-pencil tests with items that are cognitive-affective mixtures, assessments of changes in voluntary interactions with disabled persons, or reactions on projective-type tests. Measures which assessed only knowledge about the disabled did not qualify for selection, unless clearly considered by the research report author(s) to be attitude assessments; nor did measures which assessed attitudes toward mainstreaming qualify. General measures of attitudes toward children or other people were not included, unless specifically aimed at disabled persons or a particular type of disability, through instructions to the Ss or because of the context of the study—e.g., an attempt to change parents' attitudes toward their disabled children.

Measures such as sociometric scales, friendship choices, or observations of interactions were considered relevant only if clearly considered by the researcher(s) to be assessments of attitudes. Even if considered in the report to be attitude assessments, observational or other data were not included if the behaviors or responses of nondisabled Ss toward disabled persons, or the direction of behavioral or response change, could not be identified.

The search. The quest for research reports began with a computer search that included ERIC, CEC Abstracts, Dissertation Abstracts International, Index Medicus, Psychological Abstracts, and Social Science Research. The descriptor, "attitude change", was used with the broad descriptor, "disabilities", as well as with descriptors specific to types of disabilities such as "mental retardation" or "deaf". The computer search was updated

twice during the duration of the project. Hand searches of Psychological Abstracts, Education Index, and Dissertation Abstracts International were also done. Also, the references in Attitudes and Disability: An Annotated Bibliography, 1975-1981 (Regional Rehabilitation Research Institute on Attitudinal, Legal, and Leisure Barriers, George Washington University) were checked. In addition, the reference lists in all of the prior reviews cited earlier were searched, as was the reference list in each primary research report we obtained, whether or not it was decided to include the report in our review.

Copies of some 667 primary research reports that were judged potentially relevant based on title and abstract or reference in a review or primary research report were obtained through a variety of sources. The journal and the ERIC microfiche collections in the Utah State University, University of British Columbia, Simon Fraser University, and Western Washington University libraries were utilized. In addition, 218 requests for reports were sent by the Interlibrary Loan Department of the Utah State University library, of which 187 (86%) were received. Included were 77 dissertations, many of which had been identified in Dissertation Abstracts International. (No dissertation abstracts were included in the review because of the limited information they contain.) In addition, hard copies of 154 dissertations not available through Interlibrary Loan or from the authors were purchased from University Microfilms, Inc.

Each of the 667 primary research reports obtained was screened for relevance and adequacy of information. Letters were sent to authors requesting information when that in their reports was inadequate for effect size computations. One hundred and forty-six letters were sent for 117

reports. For 53 studies (45%), nothing was heard. For 13 reports, the letters were returned by the Post Office as undeliverable or someone wrote to say some such thing as that the author was dead or had moved leaving no forwarding address; for three reports, we were informed that the person to whom we wrote was not the author. For 23 reports (20%), authors wrote to tell us the information we had requested was not available. For 14 reports, information was sent that was different from that requested. Finally, for 14 reports (12%), we received information that allowed the desired effect size computations.

All told, 363 reports were discarded as irrelevant for our analysis and 31 were discarded for lack of information. (They are listed in the full research report: Shaver et al., 1987). The remaining 273 reports were the accessible population for the integrative review. (They are listed and a brief description of each study is presented in the full report, Shaver et al., 1987).

Instrumentation and Data Collection

The meta-analytic approach involves quantifying the outcomes of primary research studies using a common metric and coding various study characteristics so that it can be determined whether outcomes covary with the treatment variable and with any other study characteristics. The classification system used to code primary studies is, therefore, fundamental to data collection and data analysis. It must be comprehensive enough to "capture" the factors which are contributing to variance among studies, but not be so complex as to make coding overly burdensome. There are at least three other major considerations in developing a coding instrument: (1) That the data be collected in a usable format; (2) that the coding instrument

adequately reflect the substantive area under review; and, (3) that appropriate nontreatment study characteristics be coded.

In regard to format, a coding instrument developed at Utah State University's Early Intervention Research Institute for a meta-analysis of early intervention research with at-risk children (White & Casto, 1985) was of great value. Our prior review of research reviews helped to ensure that the second major consideration was met, as did the prior reading of a number of the primary research reports and tryouts of the instrument on research reports as it was developed. The basis for addressing the third major consideration was the literature on research design (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Shaver, 1983) and meta-analysis. Basic instrument development took place over a 3-month period; revisions continued until the scoring of new reports could be accomplished reliably, with no distortion of studies to fit the categories and no important information left out. An extensive set of conventions for coding studies was also developed.

The result of our instrument development was a coding instrument with some 162 categories, arranged in 10 sets according to the type of information to be coded, as follows: (1) General Information, such as date of publication and type of report (e.g., journal or dissertation); (2) Description of Sample, such as method of sample selection, sample size, percentage of males, educational level; (3) Treatment/Intervention, such as type of treatment (e.g., direct contact or information), the theory base, the treatment setting (e.g., classroom or mental institution), treatment characteristics (e.g., type of information and mode of information delivery), treatment verification efforts, and treatment validity; (4) Dependent Measures, such as type of measure, evidence on the reliability and validity

of scores; (5) Internal Validity, including various categories of threats, such as selection and history, and an overall rating on a three-point scale; (6) Results, including effect sizes; (7) Supplemental Information, such as whether the study was experimental or a program evaluation; (8) Prior Contact, including whether information about the subjects' prior contact with persons with disabilities was used in the analysis of data; (9) Contact (for studies of direct contact as an intervention), such as whether contact was voluntary and the relative status (e.g., education, age) of the persons involved; and, (10) Coding Summary, including who coded the study and how many minutes it took.

The quantification of results in a metric that is not relative to sample size--i.e., an effect size--is a major characteristic of meta-analytic research reviews. The major indicator of effect size for this study was Glass's Delta (Glass, McGaw, & Smith, 1981), which we labeled D. To compute a D, the difference between the experimental mean and the control group mean is divided by a standard deviation, if available, which is free of treatment effects. As our purpose was to obtain the most stable estimate of variance in the untreated population, we extended Glass's Delta by pooling the variances available for untreated groups--including treatment group pretest and control group pre- and posttest variances--to obtain the standard deviation by which the difference between means was standardized. When the means or the standard deviation for computing a D was not available, but the result from a test of significance, such as an F-ratio or t-ratio, was, D was estimated based on procedures spelled out in Glass et al. (1981).

Inter-rater Reliability

A rigorous criterion for reliability--90% agreement--was set, even though a criterion of 80% agreement is commonly used. The 90% criterion was

particularly stringent for inter-rater reliability because any categorization on which two or more of the three or four raters who were coding disagreed was coded as a disagreement.

Once adequate reliability was reached so that coding could begin, an inter-rater reliability check was conducted when any one of the raters had completed approximately 10 reports. Six separate reliability checks were completed; and for all but one, the 90% criterion was attained. For that one (85% agreement), a second study was coded, for which the criterion was met.

Because effect sizes are such a central part of a quantitative review, every effect size was re-checked for accuracy. Thirty-one errors were detected (and corrected), for an overall mean accuracy rate of 94%.

Intra-rater Reliability

After coding approximately 30 reports, each rater recoded one of the reports (selected by the project director) at the beginning of the sequence, without benefit of the first coding sheet. Again, the criterion was 90% agreement. Due to different rates of coding reports, one rater had three intra-reliability checks, one rater had two intra-reliability checks, one rater had one intra-reliability check, and one rater coded fewer than 30 reports so had no checks. All exceeded the 90% criterion.

Data Analysis

As Glass and his associates (1981, pp. 197-200) have pointed out, the role that statistical inference should play in meta-analyses is anything but clear. There was a major reason for not using inferential statistics in the integrative review reported here: the data to be analyzed constituted an accessible population, not a sample. The use of inferential statistics to

analyze data from an accessible population would be a perpetuation of ritual rather than a rationally justified procedure. Moreover, the use of an indicator of the significance of research results which is dependent upon sample size, as statistical probability is, is no more appropriate in analyzing the findings in an integrative review than it is in primary research (see, e.g., Carver, 1978; Shaver, 1985a, b).

In this study, the basic analytic approach was descriptive. Basic descriptive statistics were computed--means, modes, medians, standard deviations, and ranges. Two and three-way tables were used to investigate whether the treatment techniques and other characteristics of the studies in our accessible population were related to the size of effects.

The major analyses of da were based on the comparison of attitude change treatments against the absence of treatment--i.e., a control, placebo, or pretest condition.* When two treatment groups (i.e., Treatment A and B) were present in a study and each was compared with a control or placebo group, effect sizes were computed and coding conducted for the treatment versus control (T vs. C) or treatment versus placebo (T vs. P) comparisons, and not for the Treatment A versus Treatment B comparison. The population of studies yielded 644 T vs. C, T vs. P, and pre-post effect sizes. Some analyses were carried out on a data set of 705 effect sizes that included A vs. B comparisons.

Development of the coding instrument was guided by the admonition to include "all characteristics of the primary studies that are strongly suspected of affecting the findings . . ." (Jackson, 1978, p. 57). The

*The single-group, pre-post design is, of course, a weak form of the control group design, with the pretest serving as an indication of attitudes in a no-treatment, control situation.

upshot was a complex analysis process with difficult decisions about what to report and how. One major issue was the methodological quality of the studies in a data set.

Quality of Research

The methodological quality of the studies from which effect sizes are collected has been a source of concern since Glass (1976) renewed interest in the use of the meta-analytic approach to integrative reviews. Although the concept of analyzing for the effects of study quality is still controversial (Bangert-Drowns, 1986), our stance in planning the procedures for this review was the same as Glass's: that is, include all studies and code for quality, in order to be able to determine if effect sizes covary with study quality, rather than to include only high quality studies.

Global Quality Indicators

A number of our coding categories are related to quality of study, but three global categories are particularly appropriate indicators of methodological soundness: general treatment validity, general internal validity, and adequacy of test validity. Each is widely regarded by researchers to be central to the validity of experimental results, and each is based on information from other categories.

Summary statistics for the three global indicators of quality are presented in Table 2. Two attributes of the data are striking: First, few studies received excellent or high ratings on any of the three types of global validity. Second, none of these ratings of validity explain much of the variability in effect sizes (as indicated by the Eta^2s of .01, .02, and .03). The low correlation between quality ratings and Ds is at least in part a function of the lack of variability in the former: Few effect sizes came from studies with excellent or high ratings.

Table 2

Quality of Study Indicators

General Treatment Validity				General Internal Validity				Adequacy of Test Validity			
	Effect Sizes (Ds)				Effect Sizes (Ds)				Effect Sizes (Ds)		
Quality	N	Mean	SD	Level	N	Mean	SD	Adequacy	N	Mean	SD
Excellent	4	.25	.30	High	15	.89	.87	High	9	1.13	.69
Fair	245	.45	.68	Medium	211	.32	.58	Moderate	520	.36	.62
Poor	395	.33	.56	Low	418	.38	.61	Low	115	.40	.55
Total	644	.37	.61	Total	644	.37	.61	Total	644	.37	.61

Note. $Eta^2 = .01$

Note. $Eta^2 = .02$

Note. $Eta^2 = .02$

Specific Validity Threats

Tables 3 and 4 present data for the categories for the specific threats that formed the bases for the coders' judgments about treatment and internal validity. In coding these categories, in contrast to the global quality categories, the coder had the option of coding "Can't Tell". That was contrary to what is common in many meta-analyses in which coders are asked to make judgments about the nature of the study and any threats to validity even if study characteristics or their effects are not described in a report. Sometimes coders are instructed to assume that no threat was present if there is no evidence of the threat. Bullock and Svyantek (1985) have, however, indicated the importance of coding for missing information to indicate when adequate data were not reported to code a category. As can be seen, the availability of information necessary to decide if a threat was present varied greatly among the threats. It is also clear that forced judgments about the presence of threats would have been largely speculative for a large number of effect sizes. Essential information for determining whether threats to treatment and internal validity existed in the research is frequently missing from reports.

It is also evident in Tables 3 and 4 that there is no clear pattern of relationships between the specific indicators of treatment and internal validity and the magnitude of effect sizes. Moreover, the severity of threats to validity vary greatly from category to category. For example (Table 4), maturation was rarely a substantial, or even minor, threat to internal validity, while selection was frequently a substantial threat.

Some reviewers using quantitative techniques add up scores on individual subcategories to obtain a total quality of internal validity or methodology

Table 3

Threats to Treatment Validity for 644 Effect Sizes

	Category																	
	Can't Tell			Not Plausible			Minor			Substantial			Major			Not Applicable		
Threat	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Implementation	—	—	—	33	.47	.63	582	.38	.61	30	.18	.58	—	—	—	—	—	—
Hawthorne	160	.35	.58	93	.29	.62	234	.42	.65	157	.39	.57	—	—	—	—	—	—
John Henry	254	.40	.59	262	.36	.64	30	.22	.32	13	.11	.39	—	—	—	85	.42	.66
Treatment Diffusion	239	.36	.71	142	.29	.50	131	.46	.55	45	.35	.41	5	1.06	.87	82	.38	.62
Dissatisfaction/ Resentment	246	.31	.53	303	.39	.64	44	.53	.71	5	-.25	.53	5	.27	.25	41	.52	.74
Novelty/Disruption	183	.41	.58	382	.36	.59	52	.42	.92	7	.03	.11	1	.41	.00	19	.26	.35
Experimenter Effect/ Expectations	218	37	.56	23	.53	.63	130	.35	.70	241	.37	.59	31	.36	.76	1	.75	.00
Treatment-experimenter Confounded	82	.36	.51	37	.34	.61	98	.36	.75	221	.52	.64	205	.23	.50	1	.75	.00
Testing by treatment interaction	24	.54	.76	10	.61	.81	151	.37	.57	426	.37	.62	33	.27	.38	—	—	—
Multiple treatment interference	573	.37	.61	62	.50	.65	1	.52	.00	8	-.10	.28	—	—	—	—	—	—

Note. Means and standard deviations are for Ds.

Table 4

Threats to Internal Validity for 644 Effect Sizes

Threat	Category														
	Can't Tell			Not Plausible			Minor			Substantial			Major		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Maturation	326	.43	.63	272	.29	.58	9	.45	.45	37	.46	.65	--	--	--
History	456	.32	.55	66	.47	.69	56	.38	.66	62	.64	.79	4	.60	.36
Testing	74	.32	.50	433	.34	.60	38	.53	.56	60	.43	.56	39	.62	.92
Instrumentation	35	.58	.93	29	.29	.43	566	.36	.59	12	.32	.42	2	1.66	.23
Statistical Regression	25	.68	.84	592	.35	.59	18	.55	.60	9	1.02	.65	--	--	--
Selection	52	.38	.58	197	.41	.70	65	.36	.46	236	.38	.61	94	.28	.53
Experimental Mortality	153	.30	.46	303	.44	.66	97	.38	.70	66	.28	.56	25	.25	.41

Note. Means and standard deviations are for Ds.

score (e.g., Bullock & Svyantek, 1985). That approach was not adopted for this review, however. Summing the category scores for individual threats will not yield valid overall indicators of treatment or internal validity because any one threat may be fatal. That is, a study may be well-designed in most regards, receiving high scores on every category of internal validity but one, and thereby lack internal validity. For example, if the experimental groups were exposed to different histories that had clear potential for differential effects on outcomes, then internal validity is low regardless of how well other threats have been controlled. That presumption about treatment and internal validity is, we believe, crucial--and it is reflected in our global ratings.

Measurement Concerns

The assessment of variables involves crucial methodological judgments. For example, verification that the treatment variable was implemented as intended is crucial to the interpretation of primary research results (Cook & Campbell, 1979; Ladas, 1980; Hunter et al., 1982, pp. 95-6; Shaver, 1983). Yet for 639 of 705* effect sizes (91%), there was no report of an effort to verify treatment implementation.

Another concern with assessment has to do with the methods used to assess dependent variables--attitudes, in the case of research on modifying attitudes toward persons with disabilities. In the prior reviews, Towner (1984) in particular raised questions about attitude assessment.

---

*Although 644 T vs. P, and pre-post comparisons were the basis for our major analyses, A vs. B comparisons were included in some of our checks on methodology--increasing the N to 705 when no data were missing.

Given the concerns commonly expressed about the validity of questionnaires for predicting the behavioral aspects of attitudes, it was disconcerting to find that 66% (N = 460) of the effect sizes we obtained were based on questionnaire data (see Table 5), with items usually of the Likert-scale type. (It will not be surprising to those familiar with the research literature which we coded to know that for 44% [N = 201] of the questionnaire-based Ds, the Attitudes Toward Disabled Persons [ATDP] scale [Yuker et al., 1970; Yuker & Block, 1986] was the assessment tool.) Only 3 Ds were based on systematic observation data, and only 2 of the 4 Ds that came from telephone or mail surveys were aimed at obtaining a response to a situation removed from the research project (such as a poll of opinions toward use of university money for a Center for Disabled Students) so that Ss would not be likely to see a connection to the research and thus give biased responses (not in Table 5). It should be noted that the 39 "Other" effect sizes included those obtained with assessments of behavioral intentions (e.g., responses to questions about intent or willingness to invite a disabled person home or to volunteer to work with disabled persons). Only 12 Ds (2%) came from nonpaper-and-pencil assessments (interviews, observation, telephone surveys).

Data collection. Methods of data collection can affect study outcomes. One particularly relevant question is whether those who administered pretests and posttests were blind to the purpose of the study and to the experimental group membership of the Ss to whom they administered assessments.

For 663 out of 705 effect sizes (94%), "No" was coded for blinded collection. For 27 effect sizes (4%), blinded data collection was coded, with partial blinding (information kept from coders as to group membership or

Table 5

Frequencies of Types of Assessment
(T X C, T X P, Pre-post, A vs. B effect sizes included)

Assessment Type	Frequency
Questionnaire	460
Interview	
Structured	1
Nonstructured	2
Sociometric	10
Social Distance	63
Systematic Observation	3
Semantic Differential	79
Telephone-Mail	4
Projective Technique	2
Sentence Completion	2
Adjective Checklist	32
Rankings	4
Other	39
Total	701

whether a pre or posttest was being administered) for 5 effect sizes (nearly 1%). "Can't tell" was coded for 9 effect sizes (1%).

Test scoring is an essential part of data collection, and the blinded scoring of tests is desirable in cases where test scorers must draw inferences. As already noted, paper-and-pencil, questionnaire-type assessments were the predominant mode of dependent measure. With such assessments, coding is routine. It is not surprising, therefore, that for 674 of 705 effect sizes (96%), the category on blinded test scoring was coded as "Not Applicable". For 11 effect sizes (nearly 2%) for which blinded scoring was pertinent, it was done; for 14 effect sizes (2%), it was not.

Reliability. The reliability of the scores obtained on dependent measures is a central concern in research, as low reliability has a negative impact on validity as well as attenuating group differences. It is surprising, therefore, that for 44 percent (N = 293) of the 705 effect sizes (see Table 6), no reliability coefficient was reported. (For 9 effect sizes, the adequacy of reliability was asserted, but no coefficient was reported.) For those effect sizes for which a reliability coefficient was reported, 187 coefficients (26% of the 705 effect sizes) fell in the range of .80 to 1; 200 coefficients (28%) in the range of .60 to .79; and, 25 (3%) below .60.

Many of the researchers were apparently not mindful that reliability is an attribute of scores, not of tests, and that coefficients can vary widely by population and test administration circumstance: About 64% of the reported reliability coefficients came from other studies. (For example, when the ATDP scale was used, it was common for the authors to cite the reliability figures given in Yuker et al. [1970], and not to report a coefficient for their sample.) Another 9 percent of the reported

Table 6

Magnitude of Reliability Coefficients
for Dependent Measures in Effect Size Comparisons

	Effect Sizes	
Magnitude	N	%
None reported	293	42
.00 - .60	25	3
.60 - .79	200	28
.80 - 1.00	187	26
Total	705	99

coefficients came from pilot studies. For 4 percent of the effect sizes for which a reliability coefficient was reported, no source could be discerned. Only 24% of the coefficients reported (for 14% of the effect sizes) were computed for the samples of Ss studied.

Validity. Did test score validity fare any better than reliability in our population of studies? A crucial starting point in the consideration of the validity of measures for assessing attitudes would seem to be definition of the construct, "attitude". In our data set, no definition of "attitude" was given in 114 of 215 studies (53%), accounting for 317 of 705 effect sizes (45%) (see Table 7). For those effect sizes for which a definition was presented, a conception of attitudes as having affective, cognitive, and behavioral components was most common (N = 208; 54% of 388 effect sizes), which hardly squares with the predominance of questionnaires for attitude assessment. A definition that involved affective and cognitive components was next in frequency (N = 88; 23% of 388 effect sizes), followed by an exclusive emphasis on affect (N = 65; 17%). The other definitions (Cognitive, Behavioral, Affective and Behavioral; N = 27) constituted 7% of the effect sizes for which definitions could be coded.

Given the large proportion of effect sizes for which the object of the experimental treatment, "attitudes", was not defined, it is not surprising that for 374 out of 705 effect sizes (53%), the validity of scores on the dependent measure was not discussed.* For only 32 effect sizes (4%) was there extensive discussion of test validity. Moreover, for 94% (N = 660) of the effect sizes the dependent measure was coded as "high" in reactivity

*With attitudes clearly a psychological construct, it was perplexing that several authors (e.g., Lapp, 1974; Ozyurek, 1977) referred to the "content validity" of their attitude measure.

Table 7

Definitions of "Attitude"
in Research Reports

Type of Definition	Effect Sizes N	Effect Sizes %	Reports N	Reports %
None	317	45	114	53
Affective	65	9	22	10
Cognitive	9	1	4	2
Behavioral	5	1	3	1
Affective and Cognitive	88	12	21	10
Affective and Behavioral	13	2	3	1
Affective, Cognitive, Behavioral	208	29	48	22
Total	705	99	215	99

(with "low" [1%] and "moderate" [5%] the other choices). The adequacy of validity was rated as "moderate" for 573 effect sizes (81%), "low" for 122 (17%), and "high" for only 10 effect sizes (1%).

Time of posttest. Changes in attitudes toward persons with disabilities must be sustained, not temporary, to be of consequence. In her review, Towner (1984) called for testing to determine the "longterm effects" of treatments to modify attitudes (p. 254). Our data set confirms the need for that call. For 476 effect sizes (67%), an immediate posttest was the source of data. For 90 effect sizes (13%), the posttest was not immediate, but was delayed as much as a week to obscure the connection with the treatment. Only 89 effect sizes (13%) were based on follow-up posttesting—i.e., testing that followed an initial posttest. (For 7% of the effect sizes, the rater could not determine when the posttest was administered.)

Use of Theory

Given the scant attention paid to attitudes as a construct and to the validity of the attitude assessments used, it would have been surprising to find careful attention given to the theoretical bases for the attitude modification techniques investigated in the various studies. Only 194 effect sizes out of 705 (27%) came from comparisons in which an attitude change theory was the explicit basis for the experimental treatment. The most common basis was prior research (N = 403; 57%), with the case "well developed" for 308 effect sizes (76% of 403), with "few citations of prior studies" for 91 effect sizes (23% of 403), and with prior research "mentioned but not cited" for 4 effect sizes (1%).

As Table 8 indicates, the predominant theory either used explicitly as a base for a treatment (194 effect sizes; see paragraph above) or implicit in

Table 8

Attitude Change Theories
Underlying Experimental Treatments

Theory	Effect Sizes N	Effect Sizes %
SR, Behavioral	29	4
Conditioning	24	3
Consistency/equilibrium	518	73
Social Judgment	14	2
Functional	59	8
Combination	61	9
Total	705	99

the intervention (as judged by the rater with no direct evidence in the report; 458 effect sizes of 705, or 65%) was the consistency-equilibrium theory associated with theorists such as Festinger, Heider, Lecky, Levin, McGuire, and Newcomb. It was puzzling to find no attitude change procedures based on Rokeach's (1973) version of balance (consistency-equilibrium) theory. The data in Table 8 must be interpreted with caution, however, in light of the large number of effect sizes for which the theoretical bases for the modification technique had to be inferred. The most apt generalization is probably that the research on modifying attitudes toward disabled persons has been largely atheoretical.

Study Populations and Samples

Educational researchers often do not address in their reports the nature of their target or accessible populations, nor draw random samples from their accessible populations, make random assignments to treatments, or replicate studies to establish the stability and generalizability of results (Shaver & Norton, 1980a, b). Is that statement applicable to the body of research on modifying attitudes toward persons with disabilities?

Table 9 indicates that those doing research in this area have addressed population issues even less often than those who have published in ten years of the American Educational Research Journal (AERJ) and in two social studies journals. For a majority of the effect sizes in the reports coded for this review, there was no mention of the groups to which the authors hoped their results would be generalizable (target population--73%) or from which their samples came (accessible population--61%). In fact, few authors even used that terminology. For 7 effect sizes (1%), the term "target population" was used and the population was defined. For 3 effect sizes (.4%) the term

Table 9

Treatment of Target and Accessible Populations for 705 Effect Sizes

	Target Population				Accessible Population			
	Effect Sizes		% Reports[a]		Effect Sizes		% Reports[a]	
Category	N	%	Social Studies	AERJ	N	%	Social Studies	AERJ
Not mentioned	512	73	45	67	432	61	17	49
Term used	0	0	0	1	3	0.4	0	1
Defined	186	26	55	32	193	27	72	41
Described	0	0	0	0	0	0	1	8
Term used and Population defined	7	1	--	--	77	11	--	--
Total	705	100	100	100	705	99.4	100	100

[a]Percentages from Table 2 in Shaver and Norton (1980a), based on 53 research reports in all issues of two social studies journals through 1978 and 151 reports in the American Educational Research Journal (AERJ) for ten years, 1968-77.

"accessible population" was used, and for 77 other effect sizes (11%), the term was used and the population defined in at least rudimentary terms.

By the same token, as Table 10 shows, random sampling of individual Ss was rare. It was the means of sample selection for only 31 effect sizes (4%).* The random selection of groups provided the Ss for 31 effect sizes (4%). The use of intact groups was the most common means of obtaining a sample (N = 327 effect sizes; 46%). The use of volunteers was common (N = 237 effect sizes; 34%), and greater than for Shaver and Norton's (1980a) sample of AERJ reports (9%) and social studies reports (24%).

Table 11 presents information on assignment to groups, with the Treatment A vs. B effect sizes not included. Random assignment of the individuals or groups used as the unit of analysis was done for 35% of the effect sizes (N = 227), including 21 (3%) instances of matching followed by random assignment. This is identical to the 35% of reports of random assignment in Shaver and Norton's 10-year AERJ sample and considerably above the 9% for the reports in their social studies research sample (Shaver & Norton, 1980b).**

Replications

Related to the task of defining the populations from which samples are drawn and to which one wants to generalize is the matter of replication, as it is often argued to be the basic scientific means of establishing the reliability and generalizability of results (e.g., Shaver, 1979). As

---

*This compares to 15% and 19%, respectively, for the samples of reports from two social studies journals and AERJ reported by Shaver and Norton (1980a). The Shaver and Norton data are not reported fully in Table 10 because different categories were used.

**Information from Shaver and Norton (1980a) was not included in Table 11 because different categories were used.

Table 10

Sample Selection for
705 Effect Sizes

	Effect Sizes	
Category	N	%
Can't tell	43	6
Random--- Individuals	31	4
Random--- Groups	31	4
Volunteer	237	34
Intact Groups	327	46
Other	36	5
Total	705	99

Table 11

Assignment to Treatment Groups
for 644 Effect Sizes

Category	Effect Sizes N	%
Can't tell	23	4
Random	206	32
Match-random	21	3
Select controls randomly or matched	3	0.5
Intact groups--randomly[a]	130	20
Convenience	154	24
Other	23	4
Not applicable[b]	84	13
Total	644	100.5

[a]Intact groups assigned randomly, but not used as unit of analysis. If assigned randomly and used as unit of analysis, coded as "random".

[b]Single-group studies.

inspection of Table 12 reveals, replications have not been a common feature in studies of modifying attitudes toward disabled persons. About 1.5 percent of the effect sizes came from efforts to replicate other studies. About 12 percent of the effect sizes came from within-study replications; however, almost one-fourth of those 87 effect sizes were "quasi-replications"—effect sizes based on data gathered from different samples or in different settings in the study and coded separately even though the researchers did not recognize them as replications.

Replicability. It is noteworthy as well that for 290 of the 705 effect sizes (41%), the description of the treatment variable was not coded as adequate to allow another researcher to replicate the study. For 111 effect sizes (16%), description was coded as adequate for replication; and for 304 (43%), description was judged to be "somewhat" adequate.

A treatment must first be implemented to be replicated later. However, for only 37 effect sizes (5%) was the actual implementation of treatment judged to be complete" (Category C.8.d.). For 630 effect sizes (89%), implementation was coded as "mostly" complete, and for 38 effect sizes (5%), the treatment was rated as implemented "only in part".

Qualification of Results

The lack of high quality in the research reviewed is probably due to two factors. The first is that attitude research is difficult to conduct, especially in applied settings (e.g., in elementary schools) rather than laboratories. Another reason for the lack of high quality ratings is simply poor design and execution (as well as inadequate reporting, if better methodology was used than we were able to discern). Given the methodological deficiencies, it is important to ask to what extent the authors restricted

Table 12

Replications Among 705 Effect Sizes

Type of Replication	Effect Sizes N	Effect Sizes %
Other Research		
None	696	99
Direct	3	0.4
Systematic	6	1
Total	705	100.4
Within Study		
None	618	88
Direct	7	1
Systematic	80[a]	11
Total	705	100

[a]Includes 21 "quasi-replications"—that is, studies in which the treatment was repeated on different samples or in different settings and the results were coded separately, even though not treated as a replication by the researchers.

their conclusions in terms of the shortcomings. We coded whether conclusions were qualified by reference to sampling or design problems, possible interactions of personological or ecological variables with the experimental treatment, the assessments used, the need for replication, or "other" considerations.

Table 13 presents the results. As can be noted, for 66% of the effect sizes, the authors provided some limitation on their conclusions about the effectiveness of the technique for attitude modification. Nevertheless, the percentages of qualifications for individual shortcomings or potential concerns are low. Encouragingly, the largest percentage of qualifications (260 effect sizes, 37%) took into account combinations of factors. However, the 34% with no qualifications is an offsetting concern.

Treatment Outcomes

The results in regard to methodological quality posed a quandary. On the one hand, there appeared to be little association in our data set between the magnitude of Ds and the quality of the studies from which they come, at least as assessed via global indicators. On the other hand, it can be argued (see, e.g., Bangert-Drowns, 1986, p. 392) that unless the studies being reviewed vary widely in methodological rigor, it makes little sense to examine study quality-outcome relationships. If this review had been conducted from a stance that studies with methodological flaws should be excluded from the analysis, our data set would have shrunk appreciably.

Some might even argue that we should not have attempted any integrative review. Slavin's (1986) proposal for "best evidence" research syntheses suggests otherwise. If high quality studies do not exist, it is appropriate to "cautiously examine the less well designed studies to see if there is

Table 13

Limitations on Conclusions
About Treatment Effects
for 705 Effect Sizes

Limitation	Effect Sizes N	%
None	242	34
Sampling	100	14
Design	66	9
Measures	21	3
Interactions	5	1
Need for replication	2	0.3
Other	9	1
Combination	260	37
Total	705	99.3

adequate unbiased information to come to any conclusion" (p. 6). However, Slavin argues that a prior criteria should be applied in selecting "best evidence" studies, rather than quality-outcome analyses. We proceeded, then, with our analysis in a form of "best-evidence" review which Slavin did not intend to support. As Bangert-Drowns (1986) has pointed out, such a decision depends in large part on the purpose of the integrative review. An appropriate goal is to characterize the available research as a basis not only for insights into treatment effectiveness, but for decisions about further research. Careful summarization of the available past research is appropriate, even if only to make evident that which remains to be done. That, clearly, much remains is made even more evident by the summarization of study outcomes, which are presented briefly in this paper.

Some information from the analyses provides a context for consideration of the effect sizes for various treatments. For example (Table 14), there was nearly a balance between the number of comparisons for which the authors concluded their treatment was effective (N = 285; 44%) and those for which the treatment was deemed not to have had an effect (N = 259; 40%). Also, for 40 comparisons (6%), the results were considered by the authors to be equivocal; and, for 19 effect sizes (3%), it was concluded that the effect was negative. The actual number of effect sizes for which an attitude modification treatment group showed a negative change (that is, the treatment group's posttest mean was lower than its pretest mean) was 77 (12%), and 150 (23%) of the Ds were negative. It should not be easily assumed that the use of just any attitude modification technique will lead to a positive effect.

Comparison of Experimental Treatments

What about the outcomes of the comparisons of experimental treatment groups against control or placebo groups or pretest scores? The various

Table 14

Research Report Authors'
Conclusions *re* Treatment Effectiveness

	Effect Sizes (*Ds*)			
Conclusion	N	%	Mean	SD
None stated	42	6	.34	.41
No effect	258	40	.03	.32
Equivocal	40	6	.51	.49
Produced effect	284	44	.74	.61
Negative effect	20	3	-.63	.36
Total	644	99[a]	.37	.61

Note. $Eta^2 = .37$.

[a]On this and later tables, percentages may not always add up to 100 because of rounding error.

treatment techniques and combinations of techniques are briefly described in Table 15. They are arranged in rank order in Table 16, according to the magnitude of mean Ds. The mean effect sizes (Ds) for the attitude modification techniques can be viewed from two perspectives: (1) What does the average D for each treatment technique indicate about its effects as compared to no treatment? (2) What is indicated about the relative effectiveness of the different techniques?

Conventions to judge the magnitude of effect sizes must be used cautiously when the standards are arbitrary because there is no basis by which to judge the importance of variations in outcomes—as is the case with attitude assessments. It is, however, difficult to discuss results with no criteria in mind. Lacking more firmly grounded conventions, Cohen's (1977) criteria for small ($d = .2$), medium ($d = .5$), and large ($d = .8$) effect sizes provide a useful frame.

From that perspective, it is worth noting that none of the mean Ds reach the .8 criterion, although the mean D for the Persuasive Message studies is .67, closer to the large effect size criterion (.8) than to the medium one (.5). The differences between the Persuasive Messages mean D and the mean Ds for the other attitude modification techniques are all above an arbitrary standard for a trivial difference (.12—the magnitude of a difference between two Ds divided by the population standard deviation, .61, that would yield a $d = .2$). Moreover, in three cases, the difference is greater than the standard for a medium difference (.31), approaching the standard for a large difference (.50) in one instance.

That messages developed purposely with an argument to sway attitudes would have the largest effect size, on the average, makes sense. It also may

Table 15

Brief Descriptions of Attitude
Modification Techniques as Coded

Technique	Description
Information	Information on disabilities (e.g., etiology, characteristics, problems, similarities with nondisabled, prostheses) provided by means such as speakers, films, and books
Direct Contact	Ss in situation where they observe or interact with persons with disabilities
Vicarious Experience	Ss put in situations to help them experience what it is like to have disabilities
Persuasive Message	An argument presented via persons or printed or electronic media to convince Ss that they should have positive attitudes toward persons with disabilities
Persuasive Message, Contrast	Different messages or media used with treatment groups to investigate relative effectiveness
Systematic Desensitization	Thinking about disabled persons in relaxed, nonthreatening settings to extinguish negative attitudes
Positive Reinforcement	Use of classical or operant conditioning to modify behavior assumed to reflect attitudes
Other	Any combination of techniques other than Information Plus Direct Contact or Information Plus Vicarious Experience, which were coded separately

Table 16

Effect Sizes for Attitude Modification Techniques

Rank	Technique	Effect Sizes (Ds) N	Mean	SD	Differences Between Means[c] 2	3	4	5	6	7	8
1	Persuasive Message	23	.67	.56	.16	.24	.27	.28	.35	.38	.47
2	Information Plus Contact	100	.51	.66		.08	.11	.12	.19	.22	.31
3	Direct Contact	93	.43	.73			.03	.04	.11	.14	.23
4	Vicarious Experience	58	.40	.76				.01	.08	.11	.20
5	Other	71	.39	.64					.07	.10	.19
6	Systematic Desensitization	21	.32	.44						.03	.12
7	Information	203	.29	.51							.09
8	Information Plus Vicarious	62	.20	.36							
	Persuasive Message, Contrast[a]	11	.13	.33							
	Positive Reinforcement[b]	2	(1.74)	(.01)							
	Total	644	.37	.61							

[a]Because ten of 11 Ds came from one study, the results are considered uninterpretable and the technique is not ranked.
[b]Too few effect sizes (less than 10) to be interpretable, and so not ranked.
[c]Numbers correspond to those for ranks of techniques. For example, the difference between the Persuasive Message mean (1) and the Information Plus Contact mean (2) is .16 (.67 - .51).

be of significance that 78% of the 23 Persuasive Message effect sizes come from studies in which the theory base (S-R/behavioral for 11, congruity/equilibrium for 6, and social judgment for 6) was explicit and the relationship of the theory to the treatment well-developed. (For "explicit theory base", the closest percentage was Systematic Desensitization with 76%, dropping then to Information Plus Vicarious Experience with 31%; for "explicit relationship to treatment", the same relationship held except that "Other" was third highest, with 34%.)

The Information Plus Contact studies produced the next largest mean D, .51, just over the arbitrary criterion for a medium effect size. Note again that the Information Plus Contact mean D is .16 below that for Persuasive Messages, barely larger than the arbitrary standard for trivial differences discussed above. At the same time, the differences between Information Plus Contact, on the one hand, and Direct Contact and Vicarious Experience, on the other (.08 and .11), are both less than the .12 trivial difference standard; but the difference for the Information Plus Contact mean D equals or exceeds the .12 criterion for all other comparisons, equaling the criterion for a moderate difference (.31) in one instance.

The next three mean Ds are clustered closely together--.43 for Contact, .40 for Vicarious Experiences, and .39 for Other (combinations of techniques other than the two in Table 15)—with Ds that fall at the midpoint of Cohen's criteria for small and medium effect sizes (.2 and .5). The only difference between a mean D and one lower in the rankings that is non-trivial is between Other and Information Plus Vicarious Experience, a small difference (.19). The two remaining Ds—for Systematic Desensitization (.32) and Information (.29) are somewhat larger than the .20 small effect size standard, and only slightly higher than the means below them.

To sum up, although the mean Ds for the various techniques range from .67 to .20, clearly a broad range, there are no clear demarcations or groupings of techniques. In only one case (Persuasive Message versus Information Plus Contact) is the difference between contiguous means greater than our index of triviality (.12). The use of Persuasive Messages seems clearly to have resulted in larger Ds on the average than any other technique. Contact Plus Information runs a close second, and its use seems clearly to have produced larger Ds on the average than the use of Systematic Desensitization and the techniques ranked below it.

Treatment Variability--Heterogeneity of Ds

It might be tempting to look at the rankings in Table 16 as an index of effectiveness to be used in a singular fashion in selecting a technique to modify attitudes toward those with disabilities. That would, however, be too simplistic an interpretation of a complex set of data. To begin with, the standard deviations associated with the mean Ds serve as a reminder that the effects of each technique are not homogeneous; obviously, there is considerable overlap among the distributions of Ds for the various techniques. Moreover, it is important to remember that included in the Ds summarized by the means in Table 16 are negative values, indicating that, relative to the comparison group, a treatment had a negative rather than positive effect.

Table 17 presents a summary of the 150 negative effect sizes. Two things are worth noting: First, the percentage of negative effect sizes for each technique is roughly proportional to the percentage of effect sizes contributed to the total 644. No one technique contributed a markedly disproportionate number, or percentage, of negative Ds. But, second, it is

Table 17

Negative Effect Sizes (Ds) for the Attitude Modification Techniques

Technique	Negative Effect Sizes (Ds)				% of Negative Technique Ds[c]
	N[a]	%[b]	Mean	SD	
Persuasive Message	1/23	1/4	(-.36)[d]	(.00)[d]	(.04)[d]
Information Plus Contact	19/100	13/15	-.29	.29	19
Direct Contact	18/93	12/14	-.20	.17	19
Vicarious Experience	17/58	11/9	-.36	.42	29
Other	18/71	12/11	-.38	.31	25
Systematic Desensitization	4/21	3/3	(-.27)[d]	(.29)[d]	(19)[d]
Information	53/203	35/31	-.30	.32	26
Information Plus Vicarious	16/62	11/10	-.24	.19	26
Persuasive Message, Contrast	4/11	3/2	(-.14)[d]	(.10)[d]	(36)[d]
Positive Reinforcement	0/2	0/.3	—	—	—
Total	150/644	101/99.3	-.29	.30	23

[a]For N, the first figure is the number of negative effect sizes. The second figure is the total number of effect sizes.
[b]For %, the first figure is the percentage of the 150 negative effect sizes; the second figure is the percentage of the total 644 effect sizes.
[c]% of Negative Technique Ds is the percentage of the number of the Ds for a technique that were negative. E.g., 19% of the Information Plus Contact Ds were negative.
[d]Too few effect sizes (less than 10) to be interpretable.

remarkable that 23 percent (N = 150) of the 644 Ds were negative. In addition, for 12% of the effect sizes, the treatment group had a negative change. Those figures not only highlight the need to keep variability in mind, but raise serious questions about the adequacy of the bases for the attitude modification treatments that were investigated. It also suggests that the treatments grouped under each technique label were not necessarily alike, even though quite different from those grouped under other labels.

Concomitant Variables

A search for concomitant variables which might explain or help to make sense out of those results was not particularly fruitful. As noted above, surprisingly, "quality of study" indicators were not related to outcomes.

Treatment variation. There was a great deal of variation in treatments categorized under similar labels, such as Information and Direct Contact. For the most part, the proportion of variance in outcomes associated with these variations was low (.07 or less)--although type of experience was associated with 20% of the variance in Ds for Vicarious Experience studies, and type of message presentation was associated with 28% of the variance in Persuasive Message Ds. There were some other apparent differential effects. However, nesting of treatments within the types of disabilities toward which attitude change efforts were directed and cells that were empty, or nearly so, precluded conclusions about interactions.

Other study characteristics. Studies also differed in a variety of other ways, including the length of treatment and time of posttest, the type of dependent measures, the contexts and settings within which they were carried out, and sample size. These variations also explained very little of the variance in Ds, with no r^2 or Eta^2 greater than .05. In most cases, the

majority of effect sizes fell into one or two characteristic subcategories. For most of the variations, any relationships with D were consistent across treatments. One exception was length of treatment. The overall r for length of treatment and D was .02, but there were some differences in coefficients within treatment categories—with low negative coefficients for Information, Information Plus Contact, and Persuasive Messages, and a moderate positive r for Systematic Desensitization. Another exception was context. The predominant contexts were Elementary-Secondary Schooling and College-University, with some nesting of treatments within contexts (e.g., no Persuasive Messages or Systematic Desensitization effect sizes came from the Elementary-Secondary context) and some different results (e.g., a higher Direct Contact D in the Elementary-Secondary context, with a reversal for Vicarious Experiences). Again, nesting and empty or low N cells make difficult any conclusions about the association of treatment outcomes with other study characterisics.

Sample characteristics. Variations in sample characteristics also accounted for little of the variance in Ds, with no Eta^2 or r^2 larger than .04 for method of sample selection, grade-age level, or gender. (The relationships of prior contact and personality variables to outcomes could not be analyzed because they were basically ignored in the primary research reports.) There were some differential effects for samples selected by different methods, especially volunteers; but they were confounded with context (volunteers were more likely to come from college-university studies). There also appeared to be treatment effect size differences by grade-age levels, but with nesting and small Ns or empty cells, that could

not be discerned with certainty. There was no overall relationship between gender and outcomes, and inconsistent relations to outcomes across treatments.

Summary. As a consequence of the unevenly distributed variations, with many cells empty or with low Ns, and the nesting of treatments, the analysis of potential concomitant variables was not particularly productive, except for indicating areas to be addressed in future research. Conclusions could not be drawn about the conditions under which different attitude modification techniques had been more or less successful; rather, the major conclusion had to be that there had been a great deal of variety in the conditions under which the effectiveness of the various attitude modification techniques had been investigated, that the variations have not been systematically controlled, and that, for that reason, they confounded efforts to draw conclusions about treatment effectiveness.

Bangert-Drowns' (1986) portrayal of the general situation in summarizing psychological research provides an apt summary of the situation in regard to the variations in treatment and other study and sample characteristics as they might interact with interventions to modify attitudes toward persons with disabilities:

> Research outcomes vary in ways that make generalizable interpretations difficult. Such variation comes from a number of sources. It may reflect real population variation, the effects of different treatment features or study settings, sampling error, selection biases of the reviewer, publication biases, the effects of erroneous or insufficient reporting (unreported spurious influences, computational errors, typographical errors), differing degrees of validity and reliability in the outcome measures, and differences in the range or intensity of the independent variable. The task is enormous, but the power of social scientific inquiry would greatly increase if patterns could be found amid this outcome variation. (p. 396).

The patterns are not yet clear for the body of research we have reviewed.

What Is the Reality?

Prior reviews of the research on modifying attitudes toward persons with disabilities have not been based on comprehensive collections of research reports or on the systematic collection and analysis of extensive quantitative data on study outcomes and study characteristics. An assumption underlying this review was that the inability of prior reviewers to draw firm conclusions about the effectiveness of attitude modification techniques was likely due, at least in part, to the small samples of prior studies that were reviewed and the lack of systematic data collection and analysis. A meta-analytic type of integrative review of the research on modifying attitudes toward persons with disabilities was proposed and initiated with the hope of bringing order to the literature where other reviews had not done so. As has been made clear above, that hope turned out to be in vain. Even with a population of studies based on an exhaustive search of the literature and with a quantitative integrative review technique, clear-cut indications were not found of the overall efficacy of techniques for modifying attitudes toward disabled persons or of reliable differences in efficacy between techniques.

As a consequence of the uneven distribution among treatments of variations in sample and other study characteristics, with many cells empty or with low Ns and the nesting of treatments, the analysis of potential concomitant variables was not particularly productive, except for indicating areas to be addressed in future research. Rather than drawing conclusions about the conditions under which different attitude modification techniques had been more or less successful, the major conclusion had to be that there had been a great deal of variety in the conditions under which the

effectiveness of the various attitude modification techniques was investigated, that the variations have not been systematically controlled, and that, for that reason, they confounded efforts to draw conclusions about treatment effectiveness.

All possible data analyses could not be conducted within the time span of the funded project from which this paper has been prepared, and further analyses will be conducted for other reports to groups of professionals. However, at this time, the status of the research field might best be summarized with the flavor of a quote from Towner (1984) used earlier to indicate that another review of the literature was warranted:

> The applications [of similar techniques] yielded discouraging and contradictory findings. Both positive and negative attitudinal changes, in addition to numerous reports of [statistically] nonsignificant changes, resulted from interactions [of nondisabled persons] with disabled persons as well as from the provision of educational and general information. (p. 249)

The results of this review are likely to be disappointing for persons seeking guidelines for attitude modification programs.

Of particular importance, as emphasized in this paper, our review of research indicates the need for both better designed research and a more productive research strategy, i.e., replication, in the investigation of modifying attitudes toward persons with disabilities. However, the internal validity of attitude modification studies in this area is intrinsically frail. It is difficult to study such phenomena in applied settings, and one threat can be fatal to validity. Even with careful design and with replication, the accumulation of findings that indicate clearly what attitude modification techniques are most effective, or which are most effective with which types of persons for changing attitudes toward what types of disabilities, may turn out to be a difficult, if not impossible, goal to attain.

References

Alexander, C., & Strain, P. (1978). A review of educators' attitudes toward handicapped children and the concept of mainstreaming. Psychology in the Schools, 15, 390-396.

Anthony, W. (1972). Societal rehabilitation: Changing society's attitudes toward the physically and mentally disabled. Rehabilitation Psychology, 19, 117-126.

Bangert-Drowns, Robert L. (1986). Review of developments in meta-analytic method. Psychological Bulletin, 99(3), 388-399.

Bullock, R. J., & Svyantek, Daniel J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. Journal of Applied Psychology, 70(1), 108-115.

Campbell, Donald T., & Stanley, Julian C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Carver, Ronald P. (1978). The case against statistical significance testing. Harvard Educational Review, 48(3), 378-399.

Chubon, R. (1982). An analysis of research dealing with the attitudes of professionals toward disability. Journal of Rehabilitation, 48(1), 25-29.

Cohen, Jacob. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cook, Thomas D., & Campbell, Donald T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Chicago: Rand McNally.

Donaldson, Joy. (1980). Changing attitudes toward handicapped persons: A review and analysis of research. Exceptional Children, 46, 504-513.

Glass, Gene V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Glass, Gene V. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.

Glass, Gene V, McGaw, B., & Smith, Mary Lee. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Haddle, H. (1974). The modification of attitudes toward disabled persons: The case for using systematic desensitization as an attitude-change strategy. American Foundation for the Blind Research Bulletin, No. 28.

Harth, R. (1973). Attitudes and mental retardation: Review of the literature. Training School Bulletin, 69, 150-164.

Horne, Marcia D. (1979). Attitudes and mainstreaming: A literature review for school psychologists. Psychology in the Schools, 16, 61-67.

Horne, Marcia D. (1985). Attitudes toward handicapped students: Professional, peer, and parent reactions. Hillsdale, NJ: Lawrence Erlbaum.

Hunter, John E., Schmidt, Frank L., & Jackson, Gregg B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills: Sage.

Jackson, Gregg B. (1978). Methods for reviewing and integrating research in the social sciences. Final Technical Report to National Science Foundation for Grant #DIS 76-20398. Washington, DC: The George Washington University, Social Research Group.

Jackson, Gregg B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.

Johannsen, W. (1969). Attitudes toward mental patients: A review of empirical research. Mental Hygiene, 53, 218-228.

Ladas, H. (1980). Summarizing research: A case study. Review of Educational Research, 50(4), 597-624.

Lapp, Bernard. (1974). The effects of behavior modification inservice training on teacher behavior, teacher attitudes, and knowledge of behavior modification. Unpublished doctoral dissertation, University of Connecticut.

Ozyurek, Mehmet. (1977). Effects of live, audio, and print presentations of a discussion about physical disabilities on attitude modification toward disabled persons in Turkey. Unpublished doctoral dissertation, University of Northern Colorado.

Pulton, T. (1976). Attitudes toward the physically disabled: A review and a suggestion for producing positive attitude change. Physiotherapy Canada, 28, 83-88.

Rabkin, J. (1972). Opinions about mental illness: A review of the literature. Psychological Bulletin, 77(3), 153-171.

Rokeach, Milton. (1973). The nature of human values. New York: The Free Press.

Sandler, A., & Robinson, R. (1981). Public attitudes and community acceptance of mentally retarded persons: A review. Education and Training of the Mentally Retarded, 16, 97-103.

Segal, S. (1978). Attitudes toward the mentally ill: A review. Social Work, 23, 211-217.

Shaver, James P. (1979). The productivity of educational research and the applied-basic research distinction. Educational Researcher, 8(1), 3-9.

Shaver, James P. (1983). The verification of independent variables in teaching methods research. Educational Researcher, 12(8), 2-9.

Shaver, James P. (1985). Chance and nonsense: A conversation about interpreting tests of statistical significance, Part 1. Phi Delta Kappan, 67(Sept.), 138-141. (a)

Shaver, James P. (1985). Chance and nonsense: A conversation about interpreting tests of statistical significance, Part 2. Phi Delta Kappan, 67(Sept.), 138-141. (b)

Shaver, James P., Curtis, Charles K., Jesunathadas, Joseph, Strong, Carol J. (1987). The modification of attitudes toward persons with handicaps: A comprehensive integrative review of research. Final Report to the U. S. Department of Education, Office of Special Education and Rehabilitative Services. Project No. 023CH50160, Grant No. G008530210. Logan: Utah State University, Bureau of Research Services.

Shaver, James P., & Norton, Richard S. (1980). Populations, samples, randomness, and replication in two social studies journals. Theory and Research in Social Education, 8(2), 1-10. (a)

Shaver, James P., & Norton, Richard S. (1980). Randomness and replication in ten years of the American Educational Research Journal. Educational Researcher, 9(1), 9-15. (b)

Slavin, Robert E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. Educational Researcher, 15(9), 5-11.

Towner, Arthurlene. (1984). Modifying attitudes toward the handicapped: A review of the literature and methodology. In R. Jones (Ed.), Attitudes and attitude change in special education: Theory and practice. Reston, VA: The Council for Exceptional Children.

Triandis, Harry C., Adamopoulos, John, & Brinberg, David. (1984). Perspectives and issues in the study of attitudes. In Reginald L. Jones (Ed.), Attitudes and attitude change in special education: Theory and practice. Reston, VA: The Council for Exceptional Children.

Westwood, M., Vargo, J., & Vargo, F. (1981). Methods for promoting attitude change toward and among physically disabled persons. Journal of Applied Rehabilitation Counseling, 12(4), 220-225.

White, Karl R., & Casto, Glendon. (1985). An integrative review of early intervention efficacy studies with at-risk children: Implications for the handicapped. Analysis and Intervention in Developmental Disabilities, 5, 7-31.

Yuker, Harold E., & Block, J. R. (1986). Research with the attitudes towards disabled persons scales (ATDP): 1960-1985. Hempstead, NY: Center for the Study of Attitudes Toward Persons With Disabilities, Hofstra University.

Yuker, Harold E., Block, J. R., & Younng, Janet H. (1970). The measurement of attitudes toward disabled persons. Albertson, NY: Human Resources Center. (ED 044 853)