A Review of Behavior Analysis Research in Physical Education

Phillip Ward  
The Ohio State University  

Tim Barrett  
University of Louisville

This article provides an overview of behavior analysis, reviewing its history and the experimental research conducted in physical education settings. Articles were selected from five journals by looking through each issue to identify those that used a single-subject design to assess the effects of behavioral interventions in P–12 or teacher preparation settings. Thirty-four studies met the inclusion criteria. Studies were categorized according to their focus: (a) preservice or inservice teacher behavior; (b) student learning; (c) class management; or (d) student learning specifically focused on students with disabilities in adapted or inclusive settings. The review describes the scope of the behavioral interventions and examines the research designs used. A methodological critique suggests that while findings have been robust and the designs used were typically rigorous, researchers have not assessed generality, maintenance, or social validity as well as they might. The article closes with recommendations for reviewers and authors.

Key Words: \textbf{Need 3 or 4 Key Words not used in title}\n
Two tests of the utility and value of a science to an educational community are the extent to which its findings are (a) used as recommended practices in the preparation of teachers and (b) incorporated by teachers into everyday practice. The two tests may or may not be related, and indeed they may be accepted for different reasons. Acceptance of the former is often as much a reflection of the rigor of research as it is of the status of the science within the educational community. Acceptance of the latter is as much a reflection of what works in classrooms as it is an educational fad. This article provides an overview of behavior analysis and reports on its contributions to physical education, reviewing both its history in our field and the experimental research conducted in physical education settings.

Behavior analysis is both a basic and an applied science. Basic laboratory research is known as the experimental analysis of behavior. Applied research is known as applied behavior analysis (ABA). Both basic and applied research are grounded in the philosophy known as radical behaviorism that has its epistemological roots in the natural sciences. A clear elucidation of radical behaviorism can

Phillip Ward is with the Dept. of Sport and Exercise Education, 309 Pomerene Hall, The Ohio State University, Columbus, OH 43210-1221; Tim Barrett is with the Dept. of ??????????????, University of Louisville, Louisville, KY 40208.
be found in the collective works of Skinner (1953, 1957, 1974), and more recently Chiesa (1992) and Leigland (1992).

In their seminal paper, Baer, Wolf, and Risley (1968) suggested that ABA research should meet seven criteria: (a) it should be applied focusing on behaviors that are of importance to society and to individuals; (b) it should be behavioral, emphasizing accurate and reliable measurement of behaviors; (c) it should demonstrate analysis through the experimental manipulation of independent variables; (d) it should be technological by providing thorough descriptions of the procedures used; (e) it should be conceptually grounded in the principles of behavior analysis; (f) it should be effective, demonstrating social or practical change, not merely experimental differences between baseline and intervention; and (g) it should display generality by demonstrating effects that are sustained over time, and/or generalization to other behaviors, settings, and participants. These criteria have framed behavior analysis research for nearly half a century.

Behavior Analysis

Theory

Central to the theory of behavior analysis is the relationship between a behavior and the context in which it occurs. Context or the environment can be defined as any event that influences a person’s behavior. Thus, context includes the history that an individual brings with him or her to the moment, his/her genetic dispositions, and events in a particular setting. Such events might include circumstances that act to motivate an individual, and behaviors that are labeled as feelings and cognition. Contrary to popular belief, behavior analysis has not eschewed the existence of feelings or thinking. In fact, it has pursued a strategy to determine their contextual determinants. Moore (2000, p. 45) describes the behavior-analytic view:

private events [e.g., thinking] are behavioral phenomena rather than mental phenomena. They are not initiating causes of behavior; rather they are themselves caused by antecedent conditions, but they may contribute to discriminative control over subsequent behavior, both verbal and non-verbal.3

(emphasis added)

Thus behaviorists do not suggest that phenomena occurring within a person do not matter, nor is it their position that these phenomena cannot influence behavior. The distinction between a behavioral view and a cognitive view extends from these positions. Since behaviorists do not place any special status on private phenomena, these events can be analyzed using the same principles of behavior as those behaviors that are more easily observed (for a description of behavioral principles, see Cooper, Heron, & Heward, 1987). This analysis is central to the behavioral view on the problem of causation. Skinner (1989, p. 18) noted, “What happens inside the body is not a beginning.” With this statement he establishes the behavioral strategy of explaining causation in the context of events that lie either in the history of an individual or events in the current context that is exerting control over behavior. Thus, in terms of covert behavior such as thinking, the behavioral focus is on the question of what events occasioned thinking to begin with.

In the past 20 years behavior analysis has investigated cognitive behaviors such as rule formation, rule-governed behavior, concept formation, and choice and
decision making (Dougher & Hackbert, 2000; Wilson & Hayes, 2000). It is true, however, that the field has not pursued analysis of behaviors such as thinking with anything like the vigor it has pursued more accessible behaviors.

Behavior analysis has had its share of criticisms. Some have been the product of scholarly differences (e.g., Rodgers, 1964), but a number of criticisms (e.g., Kohn, 1993; Mahoney, 1989) have been the product of documented misrepresentation that represents a clear prejudice toward behavior analysis (for reports on the misrepresentation of behavior analysis, see Cameron & Pierce, 1994; Heward & Cooper, 1992; Reitman, 1998). For example, one criticism that has frequently been leveled at behavior analysis is that it is derived from animal research and thus has little generality to humans. Basic research in behavior analysis does use animals, as does basic research in medicine, pharmacology, and psychology. As we enter the 21st century, nearly 50 years of basic and applied research has also been conducted with human beings in educational, clinical, sport, business, and social settings that have demonstrated the effectiveness, generality, and acceptability of behavior analysis.

Reviews of the effectiveness of behavior analytic principles, interventions, and curriculums have been reported extensively in various literatures (Becker & Carnine, 1981; Cameron & Pierce, 1994; Forness, Kavale, Blum, & Lloyd, 1997; Kulik, Kulik, & Cohen, 1979). The vast majority of the studies reviewed were not conducted by behavior analysts but rather by researchers who were assessing behavioral procedures (e.g., principles, interventions and curriculums) or by researchers who had used these procedures as components of their studies. Such studies have typically employed group designs and have often been part of large-scale research projects.

To illustrate the effects of behavioral procedures, consider reviews by three independent sources. First, Walberg (1984) analyzed 3,000 studies on the effects of teacher and instructional factors on student achievement in K–12 settings. The effects of behavioral strategies such as cues, corrective feedback, and reinforcement were large, with a mean effect of one standard deviation (ranging from +.88 to +1.17). In practical terms this means that in a normal distribution, students at the 50th percentile would have their score raised to the 84th percentile.

Second, Direct Instruction (Becker & Carnine, 1981), an elementary school curriculum in language, reading, and math, has produced effect sizes that are two to four times greater than other curriculums (ABT Associates, 1977). In their national evaluation of Project Follow-Through, the independent evaluation agency ABT Associates concluded: “Critics of the model have predicted that the emphasis on tightly controlled instruction might discourage children from freely expressing themselves, and thus inhibiting the development of self-esteem and other affective skills. In fact, this is not the case” (1977, p. 73). Not only did students in the Direct Instruction model perform well on cognitive as well as affective tests, subsequent evaluation of these students indicated that they continued to perform well in middle and high school, maintaining their academic performance over their peers (Kinder & Carnine, 1991). Long-term follow-up of Direct Instruction Follow-Through students has shown that they also had higher college acceptance rates than their control peers (Gersten & Keating, 1987).

Finally, Forness et al. (1997) in their review of 18 meta-analyses of interventions of over 1,000 studies in special education reported that behavioral interventions had a mean effect size of +.93. What is impressive is not merely the size
of this statistic but rather that the next most effective treatments in special education settings were drugs, with an effect size of +.58. Providing teachers with effective pedagogical practices represents a significant social contribution by behavior analysis toward the empowerment of students with special needs.

At least two outcomes have occurred as behavior analysis has added to the empirical knowledge base of education. First, independent variables such as procedures, curriculums, or strategies have made the transition to practice. For example, examining pedagogy texts in any licensure field of education indicates that behavioral procedures and principles have successfully made the transfer from research to the knowledge bases of those fields. Second, teachers themselves have incorporated many behavioral practices into their everyday pedagogy, indicating the homogenization of behavior analysis into contemporary educational practices.

Behavior Analysis in Physical Education

The history of ABA in physical education owes much to Daryl Siedentop. In books such as Rushall and Siedentop’s (1972) *The Development and Control of Behavior in Sport and Physical Education* and Siedentop’s (1976) first and subsequent editions of *Developing Teaching Skills in Physical Education*, he provided practical guides regarding the application of behavior analysis in physical education contexts. In a series of papers, Siedentop (1981, 1982, 1983, 1984) described and proposed a research agenda that conceptually and methodologically influenced and supported the work of at least three decades of graduate students at The Ohio State University, as well as faculty and students elsewhere in the U.S. and overseas. Collectively, this work has made a distinctive contribution to the knowledge base in physical education pedagogy.

There are several examples of behavior analysis making the transition from research to practice in physical education. Behavior management principles and procedures are commonplace in pedagogy textbooks. Recently Lavay, French, and Henderson (1997), who have long been proponents of behavior analysis in adapted physical education, wrote a physical-education-specific behavior management text for teachers. Sports, Play, and Recreation for Kids (SPARK) is an elementary school physical education curriculum that uses behavioral principles explicitly throughout its content to help teachers to teach and students to meet program goals (McKenzie & Rosengard, 2000). Studies validating the effects of SPARK were large-scale, using randomized controlled groups, and are among the most rigorous studies conducted in our field (Stone, McKenzie, Welk, & Booth, 1998). The findings provide strong support for the effectiveness of behavioral procedures in physical education.

Behavioral practices have also influenced the preparation of teachers. Siedentop (1981) described the supervision program at The Ohio State University, outlining an empirical strategy designed to improve teaching preparation and reporting the results of 13 doctoral dissertations employing behavioral principles. Ocansey (1988) described a behavioral model of teacher supervision. In addition, several researchers have developed behavioral observation instruments to measure the behavior of teachers to use as criteria for the assessment of teachers. There have also been successful curriculum and workplace reform efforts based on behavioral principles (Ward, 1999).
In this section of the review we provide a summary of studies in physical education settings that used behavioral interventions and assessed their effectiveness using single-subject methodology. We report on the scope of the interventions and provide a critique of the designs used to assess them. This review is confined to research articles found in five journals that historically have been publication outlets for behavior analysis studies in physical education settings: Adapted Physical Activity Quarterly (APAQ); Journal of Applied Behavior Analysis (JABA); Journal of Behavioral Education (JoBE); Journal of Teaching in Physical Education (JTPE); and Research Quarterly for Exercise and Sport (RQES). The journals were reviewed from their first issue through to their Year 2000 volumes. Articles were selected from each journal by examining each issue to identify those that described a single-subject methodology in their methods section.

Inclusion Criteria

From the initial pool of research articles, additional criteria were applied. First, studies were included if they occurred in a physical education setting, an adapted physical education setting, or a teacher training setting where the participants were in-service teachers or preservice teachers. These criteria excluded settings such as adapted or special education one-on-one tutorial settings, camps, recreational centers, and coaching settings both in and out of school. (For a review of behavioral procedures used to teach motor skills to individuals with severe disabilities, see Demchack, 1993; for reviews of behavioral coaching interventions, see Donahue, Gillis, & King, 1980; Martin, 1992.) Also excluded from the analysis were college students participating in physical activity classes.

Coding Procedures

Once the articles were selected, each was read and coded. The following characteristics of a study were recorded: (a) the authors and date of publication; (b) the number and description of participants; (c) the setting; (d) the design; (e) the dependent variable or variables; (f) the independent variable or variables; (g) whether treatment integrity was assessed (i.e., measures indicating the intervention procedure was delivered as described); (h) whether generalization across settings, behaviors, and participants was assessed (i.e., the extent to which there were measures indicating any changes in settings, behaviors, and participants that were not part of the intervention); (i) whether maintenance was assessed (i.e., the extent to which measures were conducted to determine what happened after the intervention ended); and (j) whether social validation was assessed (i.e., how acceptable were the goals, procedures, and outcomes of the study to participants and practitioners?). Each study was then categorized according to its focus: preservice or inservice teacher behavior, student learning, class management, or student learning specifically focused on students with disabilities in adapted or inclusive settings.

Interobserver Agreement

We conducted two levels of interobserver agreement to determine the reliability of both the selection of the articles from the five journals and the reporting of the articles. To determine reliability of selection, two graduate students and a co-author who were not involved in the original selection of the papers indepen-
dently selected from the journals those articles they determined met the selection criteria. These selections were compared to the original selection conducted by the first author. On two occasions the selections differed. Both involved studies reported in the journal APAQ and were differences related to the definition of the setting used in the selection process. The differences were resolved upon a closer reading of the settings in the two studies, and it was determined that the two studies did not meet the selection criteria.

A two-step process was used to report the studies. First, the characteristics of each study described in the coding procedures were entered into a table. Each characteristic was reported as a cell in the table by each author independent of the other. Both tables were compared to determine the extent of agreement between them. The percentage of interobserver agreement (IOA) between the two tables was calculated using the following formula: cell agreement divided by the number of agreements plus disagreements multiplied by 100. The percentage of IOA agreement was 98.5. Second, disagreements were identified. Disagreements occurred over the description of participants and description of an independent variable. Both authors, with reference to the relevant article, then came to a consensus as to how the information in the cell was to be reported.

Results and Discussion

A total of 34 studies met the inclusion criteria. The largest number of papers was found in JTPE (n = 19), followed by APAQ (6), RQES (4), JoBE (3), and JABA (2). A considerable number of studies did not meet criteria. Most were excluded because they were not conducted in naturalistic settings with intact classes. Figure 1 presents a cumulative record of the number of studies published, beginning in 1976 and ending in 2000. Thus the studies included in this review span a 24-year period. Analysis by decade (see Figure 1) indicates that one study was reported in 1970–79, 13 studies in 1980–89, and 19 studies in 1990–99. In the following

![Figure 1 — Cumulative record of the number of studies that met the inclusion crite-](image-url)
analysis the studies are described according to their focus, then methodological considerations are discussed.

**Focus of Studies**

Each study was grouped into one of four categories consisting of those studies that focused on teacher training, class management, student learning, and instruction of students with disabilities. Some studies could have been placed in more than one category, such as those focused on management but which also included students with disabilities as participants, and studies that focused on teacher training but used student process variables as measures of effectiveness.

**Studies Focusing on Teacher Training Interventions.** Table 1 summarizes the characteristics of 18 studies that were designed to improve a preservice or inservice teacher’s behavior. This was the largest category of behavior analytic research. The majority of studies focused on preservice teachers; however, five studies used inservice teachers as participants. Independent variables included (a) variations of systematic feedback such as verbal (e.g., Paese, 1984), self-monitoring (e.g., Eldar, 1990), use of videotape (e.g., Sharpe, Spies, Newman, & Spickelmier-Vallin, 1996), and audiocueing (e.g., van der Mars, 1987); (b) directed rehearsal, a positive-practice procedure (e.g., Ward, Smith, & Makasci, 1997); and (c) supervisory approaches such as the use of principals as change agents (Ratliffe, 1986), behavioral supervision models (e.g., Ocansey, 1988), and competency-based training packages (e.g., Darst, 1976). Dependent variables were defined in terms of teacher and student process variables, but have also included the conferencing behaviors of cooperating teachers. Most studies used student behaviors as measures of the effectiveness of the intervention.

The studies in this section contribute to a technical orientation of teacher preparation and stress pedagogical knowledge (O’Sullivan, 1996). Two studies from this section warrant further comment. First, Wilkinson’s (1991) training program to help preservice teachers improve their qualitative analysis skills represents an understudied area in physical education. Her use of stimulus generalization to promote transfer of learning in volleyball, together with her later work (Wilkinson, 1996), represents an important technology to help teachers analyze correct skill performance. Second, Rolider, Siedentop, and van Houten (1984) reported on their success in defining, training, and validating enthusiastic teaching behavior with a group of 15 preservice teachers. Their study represents one of the few studies we reviewed that examined behaviors in the affective domain (see instructional focus section for others). Studies focusing on the affective domain are particularly important for two reasons. First, there is a strong relationship between student outcomes and the frequency and quality of teacher and student social interactions (Wang, Haertel, & Walberg, 1993). Second, studies that improve teacher and student social skills also affect the quality of life in schools for teachers and students (Doutis & Ward, 1999).

**Studies Focusing on Class Management Interventions.** Table 2 summarizes the four studies addressing student or class management in physical education. We found the low number of studies surprising, given that behavior analysis is most often associated with class and student management procedures. The most recent review we could find of class management issues in physical education was conducted by Luke (1989), who cited fewer than 10 physical education specific
Table 1  Studies Focusing on Teacher Training Interventions in Physical Education Settings

<table>
<thead>
<tr>
<th>Studies</th>
<th>Participants</th>
<th>Settinga</th>
<th>Design</th>
<th>Dependent variables</th>
<th>Independent variablesb</th>
<th>Trmt. integrity?</th>
<th>Genaliz. assessed?</th>
<th>Mainten. assessed?</th>
<th>Soc. valid assessed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Darst (1976)</td>
<td>7 preservice teachers (PT) (3M, 4F)</td>
<td>Elementary internship</td>
<td>Mult. baseline design across teachers</td>
<td>Teacher feedback (FB)</td>
<td>Competency-based package of instruction, &amp; graphic FB, goal setting</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Darst &amp; Steeves (1980)</td>
<td>7 preservice teachers (4 M, 3 F)</td>
<td>Secondary internship</td>
<td>Mult. baseline design across teachers</td>
<td>Teacher FB, student active &amp; appro. behavior</td>
<td>Same as Darst (1976)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Paese (1984)</td>
<td>3 female preservice teachers</td>
<td>Internship Grades 1 or 2</td>
<td>Mult. baseline design across teachers w/2 interventions &amp; post checks</td>
<td>Teacher’s verbal interactions</td>
<td>(A) Baseline; (B) Feedback; (C) Self-monitoring</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Rilder, Siedentop, &amp; van Houten (1984)</td>
<td>15 preservice teachers (5 M, 10 F) in 3 groups of 6, 4, &amp; 5 teachers</td>
<td>Internship Grades 3–12</td>
<td>Mult. baseline design across groups of teachers</td>
<td>Teacher’s enthusiasm, student activity data probe</td>
<td>Enthusiasm training program, &amp; daily FB</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Dunbar &amp; O’Sullivan (1986)</td>
<td>2 female classroom teachers who taught Au: taught what?!\</td>
<td>PE taught by classroom generalists Grades 1 &amp; 2</td>
<td>Mult. baseline design across behaviors</td>
<td>Teacher FB, questioning, &amp; use of student demonstrators</td>
<td>Verbal &amp; graphic feedback</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ratcliffe (1986)</td>
<td>2 principals (both F) and 2 teachers</td>
<td>One parochial and one public</td>
<td>Mult. baseline design across</td>
<td>Student mgmt. time, student</td>
<td>(A) Baseline; (B) Principal stan-</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

(Continued)
Table 1 (Cont.)  Studies Focusing on Teacher Training Interventions in Physical Education Settings

<table>
<thead>
<tr>
<th>Studies</th>
<th>Participants</th>
<th>Settinga</th>
<th>Design</th>
<th>Dependent variables</th>
<th>Independent variablesb</th>
<th>Tmt. integrity?</th>
<th>Genaliz.</th>
<th>Mainten.</th>
<th>Soc. validity assessed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>van der Mars (1987)</td>
<td>1 female preservice teacher</td>
<td>Internship K–3 grades</td>
<td>A-B-A-B reversal design</td>
<td>Teacher praise, &amp; feedback</td>
<td>Audio-cueing</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>van der Mars (1988)</td>
<td>1 male inservice teacher</td>
<td>2nd grade</td>
<td>Mult. baseline design across behaviors</td>
<td>Teacher praise, feedback, &amp; mgmt time</td>
<td>Audio-cueing</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ocansey (1988)</td>
<td>4 cooperating teachers [gender not reported]</td>
<td>2 elementary &amp; 2 secondary classes</td>
<td>Mult. baseline design across cooperating teachers</td>
<td>Cooperating teachers planning, explicit task stmts, accountability stmts</td>
<td>Behavioral model of supervision</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Eldar (1990)</td>
<td>39 preservice teachers (20 M, 19 F)</td>
<td>Secondary internship</td>
<td>Mult. baseline design across behaviors w/ reversal</td>
<td>Teacher &amp; student process variables (varied by teacher)</td>
<td>Self-mgmt. program</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Grant, Ballard, &amp; Glynn (1990)</td>
<td>3 inservice teachers (2 M, 1 F) in 3 schools, 2 students randomly selected from ea. class</td>
<td>Grade 8 classes, volleyball unit</td>
<td>Mult. baseline design across teachers</td>
<td>Student motor-on-task behavior &amp; learning trials</td>
<td>FB on student performance, &amp; lesson time</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Wilkinson (1991)</td>
<td>6 PTs (2 high scorers, 2 medium scorers, 2 low scorers)</td>
<td>Univ. teacher educ. majors, volleyball skill</td>
<td>Mult. baseline design across behaviors</td>
<td>Analyzed critical features of forearm pass, overhead pass,</td>
<td>Videotaped visual information training</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Study</td>
<td>Participants</td>
<td>Class</td>
<td>Program</td>
<td>Options</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>--------------</td>
<td>-------</td>
<td>---------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sariscany, Darst, &amp; van der Mars (1995)</td>
<td>3 inservice teachers (2 M, 1 F) and 3 male students identified as inattentive &amp; off-task</td>
<td>Junior high volleyball unit</td>
<td>Alternating treatments design</td>
<td>Student on-task behavior, approp. skill trials of underhand serve &amp; forearm pass</td>
<td>(A) Close superv. w/FB; (B) Distant supervision w/FB; (C) Distant superv. w/ no FB</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sharpe, Spies, Newman, &amp; Spickelmier-Vallin (1996)</td>
<td>7 inservice teachers (6 M, 1 F)</td>
<td>Middle school team sports &amp; fitness units</td>
<td>Multi-treatment reversal design A-B-A-C</td>
<td>Class time &amp; teacher process variables, &amp; self-monitoring accuracy</td>
<td>(A) Baseline; (B) Verbal-only FB; (C) Verbal + videotape FB</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ward, Smith, &amp; Makasci (1997)</td>
<td>6 preservice teachers (5 M, 1 F)</td>
<td>Elementary internship K–5</td>
<td>Mult. baseline design across behaviors</td>
<td>Teacher lesson intro, task presentations, subject matter FB, &amp; lesson closures</td>
<td>Directed rehearsal</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Sharpe, Newman, &amp; Lounsbery, &amp; Bahls (1997)</td>
<td>4 preservice teachers (2 M, 1 F)</td>
<td>Internship</td>
<td>Mult. baseline design across students</td>
<td>Teacher instruct. actions, class time off-task, engaged in activity, &amp; lesson organization</td>
<td>behavior chains associated with student practice</td>
<td>FB on sequential</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ward, Johnson, &amp; Konukman (1998)</td>
<td>4 preservice teachers (3 M, 1 F)</td>
<td>Elementary internship K–5</td>
<td>Mult. baseline design across behaviors</td>
<td>Student transitions, teacher lesson intro, &amp; task presentations</td>
<td>Directed rehearsal</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Lounsbery &amp; Sharpe (1999)</td>
<td>4 preservice teachers (3 M, 1 F)</td>
<td>Secondary internship 6–12</td>
<td>A-B design w/ maintenance phase</td>
<td>Teacher behaviors, interactions, &amp; conditional probabilities of student skill practice</td>
<td>FB on sequential behavior chains associated with student practice</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

*For simplicity, “internship” refers to field-based preservice teaching experiences with participants teaching physical education. “Student” refers to pupils in the class. Preservice or inservice teachers are described as such. Notations consistent with Cooper et al. (1987) are used. “A” refers to a baseline or beginning experimental condition, other letters refer to independent variables. The conditions are symbolized by capital letters in parentheses.*
### Table 2  Studies Focusing on Class Management Interventions in Physical Education Settings

<table>
<thead>
<tr>
<th>Studies</th>
<th>Participants</th>
<th>Setting</th>
<th>Design</th>
<th>Dependent variables</th>
<th>Independent variables&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Treatment integrity?</th>
<th>Generaliz. Mainten.</th>
<th>Soc. validity assessed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paese (1982)</td>
<td>113 students (60 M, 53 F) in 3 classes</td>
<td>High school volleyball unit</td>
<td>Mult. baseline design across classes</td>
<td>Students who “dressed out” for PE</td>
<td>Interdependent group contingency system</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Vogler &amp; French (1983)</td>
<td>23 students (10 M, 2 F) ages 6–9, 11 students (10 M, 1 F) ages 10–12</td>
<td>Segregated PE, 2 self-contained special ed classrooms</td>
<td>A-B-A-B reversal design</td>
<td>On-task behavior</td>
<td>Individual contingency for off-task behavior</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Jeltma &amp; Vogler (1985)</td>
<td>9 students (8 M, 1 F) ages 8–10, 4 (all boys) age 11–13</td>
<td>2 segregated PE classes from self-contained special ed classrooms</td>
<td>A-B-A-B reversal design</td>
<td>On-task behavior</td>
<td>Group &amp; indiv. contingencies &amp; behavioral contracts</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>White &amp; Bailey (1990)</td>
<td>2 classes: one w/30 students, &amp; alternative 4th &amp; 5th grade class w/14 boys w/severe behavior problems</td>
<td>Grades 4 &amp; 5 PE</td>
<td>Mult. baseline design across classes</td>
<td>Disruptive behaviors</td>
<td>3-minute sit &amp; watch strategy, contingent upon disruptive behavior</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<sup>a</sup>Notations consistent with those of Cooper et al. (1987) are used. “A” refers to a baseline or beginning experimental condition, other letters of the alphabet refer to independent variables.
papers and only 5 of those were research studies. The studies in Table 2 were all conducted in the 1980s, with one exception published in 1990, which perhaps reflects the move away from management concerns to other foci reported in this review. The interventions were variations of group contingencies (Jeltma & Vogler, 1985; Paese, 1982) and individual contingencies (Vogler & French, 1983; White & Bailey, 1990). Two studies were conducted in regular physical classes and two in segregated adapted physical education classes.

**Studies Focusing on Instruction.** Table 3 summarizes six studies focused on instructional strategies in physical education. Interventions included a sportsmanship curriculum (Sharpe, Brown, & Cramer, 1995), peer-mediated accountability (e.g., Ward, Smith, Makasci, & Crouch, 1998), specific teaching skills (Sharpe, Hawkins, & Wiegand, 1989), and the good behavior game (Patrick, Ward, & Crouch, 1998). In most cases, studies compared the instructional intervention to instruction occurring in the setting that served as a baseline. Dependent variables included social skills such as appropriate and inappropriate fair play behaviors, and student process variables such as correct trials defined in terms of the presence or absence of critical elements.

The studies were conducted exclusively in elementary schools, 2nd through 6th grade, which reflects an overall trend in instructional interventions in physical education that were conducted in elementary settings compared to secondary settings. Interestingly, even though single-subject designs were employed, participants in some of these studies included all students in a class and thus the numbers of participants were quite large. Three of the studies reported participant numbers ranging from 67 to 85 students (Crouch, Ward, & Patrick, 1997; Patrick et al., 1998; Sharpe et al., 1995). Typically in single-subject designs the students are selected randomly, purposively (e.g., high skilled and low skilled students), or based on convenience (i.e., students who are easy to identify). In studies with larger numbers of participants, data were collected on all class members and reported as a total measure. While there are major disadvantages to presenting data this way, for example it obscures the behaviors of individuals, there are also advantages when the experimental question addresses the impact of an intervention in a particular setting.

**Studies Focusing on Participants with Disabilities.** Table 4 reports the characteristics of six studies that focused explicitly on participants with disabilities. These studies were conducted in both segregated and integrated settings. Three studies assessed the use of typical functioning peers as tutors for students with disabilities (Houston-Wilson, Dunn, van der Mars, & McCubbin, 1997; Lieberman, Dunn, van der Mars, & McCubbin, 2000; Webster, 1987). One study assessed the effects of an integrated versus a segregated setting on preschool students’ motor performance (Zittel & McCubbin, 1996). Other studies focused on teaching students with disabilities how to self-pace their aerobic exercise (Ellis, Cress, & Spellman, 1993), and another assessed the efficacy of an instructional method (Yang & Porretta, 1999).

Studies included students with behavioral disorders, students with mental retardation, and also students with hearing impairments. These studies were conducted in segregated settings where physical education occurred in the context of special education or adapted physical education; in integrated settings where
<table>
<thead>
<tr>
<th>Studies</th>
<th>Participants</th>
<th>Setting</th>
<th>Design</th>
<th>Dependent variables</th>
<th>Independent variables</th>
<th>Trmt integrity?</th>
<th>Generaliz. assessed?</th>
<th>Mainten. assessed?</th>
<th>Soc. valid assessed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharpe, Hawkins, &amp; Wiegand (1989)</td>
<td>8 students (4 M, 4 F)</td>
<td>Two 4th-grade basketball units</td>
<td>A-B design w/follow-up</td>
<td>ALT-PE variables (e.g., off-task, on-task, motor appropriate)</td>
<td>Model/practice &amp; verbal/rehearsal systems skills</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Sharpe, Brown, &amp; Crider (1995)</td>
<td>3 classes (1 was control) 85 students (52% M, 48% F)</td>
<td>Three 3rd-grade team sports units &amp; classroom behavior</td>
<td>Multi. baseline design across classes combined w/mult. treatments.</td>
<td>Conflicts, leadership behaviors, resolutions &amp; off-task behav. Skill engagement &amp; lesson organiz.</td>
<td>(A) Baseline; (B) Sportsmanship curriculum; (C) Nonverbal prompting</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Crouch, Ward, &amp; Patrick (1997)</td>
<td>3 classes, 67 students (54% M, 46% F)</td>
<td>Grades 4–6 volleyball units</td>
<td>Multi-treatment reversal design w/ conditions yoked across classes</td>
<td>Learning trials</td>
<td>(A) Baseline; (B) Peer dyads; (C) Peer-mediated accountability</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Patrick, Ward, &amp; Crouch (1988)</td>
<td>3 classes, 67 students (50% M, 50% F)</td>
<td>Grades 4–6 volleyball units</td>
<td>Multi. baseline design across classes</td>
<td>Appropriate &amp; inappropriate social, learning trials</td>
<td>Modified version of the good behavior game</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ward, Crouch, &amp; Patrick (1998)</td>
<td>24 students (14 M, 10 F)</td>
<td>4th-grade volleyball units</td>
<td>Multi-treatment/reversal design</td>
<td>Learning trials</td>
<td>Peer-mediated accountability</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ward, Smith, Makasci, &amp; Crouch (1998)</td>
<td>9 students (3 M, 6 F)</td>
<td>Grades 4–5 basketball units</td>
<td>Multi. baseline design across classes/students</td>
<td>Learning trials</td>
<td>Peer-mediated accountability</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

*Notations consistent with Cooper et al. (1987) are used. “A” refers to a baseline or beginning experimental condition, other letters refer to independent variables. The conditions are symbolized by capital letters in parentheses.*
students transitioned from a special education home room into a class of typical functioning peers for the physical education lessons with or without teacher aides; and in inclusive settings where students with disabilities were members of an intact class.

**Methodological Considerations**

*Designs*

The most common design reported in this review was the multiple baseline design, or MBD ($n = 20$). This design uses a time-lagged strategy to assess internal validity when changes in the data path (level and trend) plotted on the first tier occur at the point of intervention, without changes occurring in the underlying tiers. This effect, when reproduced in Tiers 2 and 3, increases confidence that changes in dependent variables are due to the presence of the independent variable. The MBD can be used to assess the effects on an independent variable across behaviors (e.g., motor skills and teaching skills), settings (e.g., inclusive vs. segregated), and participants (e.g., teachers or students). The most common use was across participants ($n = 10$), though several studies used the MBD across behaviors ($n = 6$) and settings ($n = 4$).

One reason for the popularity of the MBD is that it is the design of choice when dealing with behaviors that, once acquired, remain in the individual’s repertoire (e.g., learning). Most other single-subject designs require that treatment be withdrawn in order to demonstrate experimental control. The MBD does not require that an intervention be reversed, thus it is ideally suited for showing causal relationships when the dependent variable is a direct or proxy measure of learning.

Another popular design ($n = 8$) was the reversal design (e.g., A-B-A-B). In this design a baseline is established and then an independent variable is introduced, then removed and then reintroduced. Experimental control is demonstrated when changes in the dependent variable occur as a function of presence and absence of the independent variable. Several of the studies incorporated the reversal design with multiple treatments (e.g., A-B-A-C).

The reversal design is a good choice when researchers wish to demonstrate the effects of one or more treatments against a baseline. Unlike the MBD, the reversal design requires that a behavior be under strong control of a baseline and/or an independent variable, thus being “reversible.” For example, in studies using the peer-mediated accountability variable, changes in the level of success and the total number of trials have been demonstrated using reversal designs. From a theoretical perspective, peer-mediated accountability did not intervene on learning, but rather on a set of motivating conditions called establishing operations. Establishing operations are environmental events that increase the reinforcing effectiveness of other events (Michael, 1993). In the case of peer-mediated accountability, the intervention exerts an influence over the performance of students that results in an improved number of trials and often increases successful performance. This effect is identical to the basketball coach indicating 45 seconds left on the shot clock and the resultant performances that are evoked by that condition. In these circumstances, reversal designs are the designs of choice.

Other designs that were used include the A-B design and the delayed multiple-baseline design, which are inherently weak in demonstrating experimental control but are sometimes used because the naturalistic setting is not amenable to
<table>
<thead>
<tr>
<th>Studies</th>
<th>Participants</th>
<th>Setting</th>
<th>Design</th>
<th>Dependent variables</th>
<th>Independent variables $^a$</th>
<th>Trmt integrity?</th>
<th>General. Mainten.</th>
<th>Soc. valid. assessed?</th>
<th>assessed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Webster (1987)</td>
<td>3 students:</td>
<td>Adapted K–12 PE</td>
<td>Mult. baseline design across students w/multi-treatment design (A-B-A-C-A)</td>
<td>Motor approp. category of ALT-PE instrument</td>
<td>(A) No tutors; (B) Untrained tutors; (C) Trained</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>M, age 19, w/mod. mental retardation; M, age 21, F, age 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ellis, Cress, &amp; Spellman (1993)</td>
<td>Study 1: 6 students (5 M, 1 F) ages 12–18 yrs, w/ mental retardation</td>
<td>Adapted PE</td>
<td>Study 1: Reversal design A-B-A-B</td>
<td>Studies 1&amp;2: Heart rate measured while on exercycles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Study 2: 5 students (3 M, 2 F), ages 18–19, w/mental retardation</td>
<td></td>
<td>Study 2: Changing criterion design w/baseline &amp; w/follow-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zittel &amp;</td>
<td>3 boys, 1 girl</td>
<td>Preschool PE</td>
<td>Reversal design</td>
<td>Critical elements</td>
<td>(A) Segregated</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Study</td>
<td>Participants</td>
<td>Setting</td>
<td>Condition Details</td>
<td>Outcome Measures</td>
<td>Peer Tutoring</td>
<td>Untrained Tutoring</td>
<td>Comments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>----------------------------</td>
<td>---------------</td>
<td>--------------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McCubbin (1996)</td>
<td>ages 3–5 yrs, developmental delays</td>
<td>in integrated &amp; segregated settings</td>
<td>A-B-A-B present during trials of locomotor &amp; object control skills</td>
<td>PE setting;</td>
<td>Yes</td>
<td>No</td>
<td>(B) Condition integrated PE setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Houston-Wilson, Dunn, van der Mars, &amp; McCubbin (1997)</td>
<td>5 boys, 1 girl ages 9–11 yrs, w/mild mental retardation, paired w/6 typically developing peers</td>
<td>Integrated PE (mainstreamed) setting</td>
<td>Delayed multiple baseline design across students w/2 experimental conditions</td>
<td>Horizontal jump, catch, overhand throw, forehead strike, &amp; sidearm strike</td>
<td></td>
<td></td>
<td>(A) Baseline target students alone, untrained peer tutors; (B) Trained peer tutors; (C) condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yang &amp; Porretta (1999)</td>
<td>6 students (3 M, 3 F) ages 16–18, w/mild mental retardation</td>
<td>Adapted PE setting</td>
<td>Multiple baseline design across skills</td>
<td>Sports skills</td>
<td>Yes</td>
<td>Yes</td>
<td>4-step learning strategy (ready, look, do, score)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lieberman, Dunn, van der Mars, &amp; McCubbin (2000)</td>
<td>8 deaf students (4 M, 4 F), who were paired w/8 hearing students</td>
<td>4th–6th grade inclusive PE classes</td>
<td>Delayed multiple baseline design across students</td>
<td>MVPA measured using SOFIT</td>
<td></td>
<td>No</td>
<td>Peer tutoring</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Notations consistent with Cooper et al. (1987) are used. “A” refers to a baseline or beginning experimental condition, other letters refer to independent variables. The conditions are symbolized by capitals letters in parentheses.*
a study requiring an extended duration or continued experimental manipulation of the independent variable. Four studies combined both the MBD and the reversal design. Also used was the alternating treatment and the changing criterion designs, which are among the more rigorous single-subject research.

Experimental Effects

There are three issues related to the experiments reported in the reviewed articles that warrant discussion: (a) effects that suggest the influence of rule-governed behavior; (b) effects that produced stable intervention levels from variable baselines; and (c) the extent to which studies reported treatment integrity.

Effects That Suggest the Influence of Rule-Governed Behavior. Examination of the effects of independent variables in the studies reviewed indicates that the changes were robust with two general patterns: (a) changes in the level of the dependent variables occurring immediately following the presentation of the intervention, before individuals could have come into contact with the consequences of the intervention (e.g., next-day feedback on teaching performance); and (b) changes in trend that occurred immediately, but which show a pattern of improvement over time. These effects represent two different behavioral processes. First, when changes occur prior to contact with a consequence, the behaviors are said to be rule-governed (Skinner, 1974). If a behavior is rule-governed, changes in the behavior occur because the participant followed some instruction, and following instructions has been reinforced in the past. Second, if changes in the dependent variable occur less dramatically following the intervention, but are discernible as a change in trend, then such behavior is often described as contingency-shaped behavior (Skinner, 1974).

There are at least two types of rule-governed behavior—pliance and tracking (Hayes, Zettle, & Rosenfarb, 1989). Pliance, derived from compliance, occurs when following rules such as instructions from a researcher or university instructor are mediated by persons who also know the rule (e.g., a cooperating teacher, university supervisor, or researcher) and those individuals have some form of formal or informal record of the behavior. This contingency requires a history of following instructions in other settings (e.g., P–12 schooling, home, sports teams). Thus, one explanation for the immediacy of change in many of the teacher training studies reviewed is pliance.

A second explanation for rule-governed effects in these studies is tracking (Hayes et al., 1989). Tracking occurs when instructions place the individual in contact with natural environmental consequences. This commonly occurs when the instruction describes the natural consequences. For example, when a preservice teacher follows directions such as “You could reduce lineups by using multiple equipment distribution points,” he or she comes into contact with the consequences of efficient equipment distribution.

The distinction between pliance and tracking is important in determining experimental relationships. In pliance the behavior change is under the control of the rule-giver, whereas in tracking the behavior change is at least in part under the control of the natural consequences of behaving in the stated manner. Over time, natural contingencies must take control of pliance-developed behavior if it is to be maintained. This occurs as teachers increasingly come into contact with the
consequences of their teaching. Discovering conditions in teacher training that use tracking contingencies may help prevent the washout effects reported by Zeichner and Tabacnick (1981).

This discussion of rule-governed behavior illustrates what we see as a shortcoming of research to date, namely the absence of theory. Many behavioral researchers in education have focused on pragmatic behavior change, relying on the strong basic and applied programmatic research history in behavior analysis to support findings. There is little reference to theory in the vast majority of studies we reviewed. Yet the focus on explanations of behavior has been the central feature of behavior analysis for at least 50 years (Heward & Cooper, 1992).

Effects That Produced Stable Intervention Levels From Variable Baselines. Asher common effect of interventions was to take behavior that was variable during baseline and make it stable at a different level. This was particularly the case with preservice teacher behaviors. We suspect that several of the teaching behaviors reported during the baseline were variable because the teachers could not discriminate clearly the conditions in which to use them. Thus the effect of the intervention was to clarify for the teacher the correct contextual and procedural use of the behavior.

This poses a design problem for the researcher. It is an axiom of single-subject research methodology that the independent variable should not be introduced until baselines are stable (Cooper et al., 1987). As reported, many of the baselines, particularly in studies reviewed in Table 1, were variable at the time of intervention. While intervening on a variable baseline certainly weakens the internal validity of a study, there are practical considerations that mitigate in favor of intervention on a variable baseline. Some of these include the length of time in the setting, the extent to which researchers are prepared to allow undesirable or ineffective teaching behaviors to continue, and conditions in which the presence of the investigators is reactive. While researchers should strive to maintain the highest rigor, and conducting studies in fields settings is not an excuse to abandon rigor, it is often true that in order to maintain ecological validity, some experimental rigor is lost. In contrast, one of the strongest characteristics of the research reported in this review has been the ecological validity of the studies.

Treatment Integrity. Treatment integrity describes the extent to which the intervention is implemented as described in the methods section (Cooper et al., 1987). Treatment integrity provides assurance that no additional or unplanned variables were included in the intervention. Measures of the reliability, often called procedural reliability, are reported in a similar manner as IOA measures. In this review, only 7 of the 34 studies (20%) reported a measure of treatment integrity.

The lack of treatment integrity measures is not new, nor is it confined to physical education. Indeed, the criticism can be fairly leveled at most experimental studies reported in educational and psychological journals. Peterson, Homer, and Wonderlich (1982) describe the problem as a double standard whereby scientific communities require high standards for the specification and measurement of the dependent variables, including measures of reliability, but require little evidence of the specification and measurement of the independent variable. Considering the importance placed on internal validity in single-subject methodology, we consider this an area of procedure in need of immediate improvement.
Assessment of Generalization, Maintenance, and Social Validity

Given the significance of generalization, maintenance, and social validity in applied behavior analysis, less than 50% of studies performed one or more of these assessments. One exemplary study demonstrating all three strategies was by Yang and Porretta (1999). Studies in the 1990s were more likely to conduct one or more of these assessments than those undertaken in earlier decades. Nonetheless, generalization of the intervention to other settings, behaviors, or participants was assessed in only 5 of 34 (15%) studies reviewed. The common strategy used to assess response generality is to concurrently assess the effects on an intervention in nonintervention conditions. This is an important consideration in experimental research. For example, the multiple baseline design across behaviors was used in several of the teacher training studies we reviewed.

The experimental logic of this design relies heavily on the fact that changes in the behaviors that are untreated do not occur when other behaviors are intervened upon. If generality were relatively easy to produce, then we might expect that the intervention could affect these untreated behaviors which also occur in the teaching/learning setting. This effect is called induction. An examination of the graphs in these studies does not reveal an induction effect. This finding shows that these teaching behaviors are very much under strong stimulus control in both baseline and intervention. Because of this finding, we suggest that programming for generality should be a design consideration for researchers in physical education (see Cooper et al., 1987, for a description of programming strategies for generality).

Maintenance measures are used to determine the extent to which the effects of the intervention are maintained after the intervention is concluded. Maintenance, or more correctly, response maintenance, is one dimension of generalization. Because of the importance attached to the durability of outcomes over time, we chose to report on it separately. Maintenance was assessed in 12 (35%) of the studies reviewed. The studies report some measures taken as soon as 1 week following the end of the study and as late as 6 months after the study. The most common measure was within a month of the conclusion of the study. Maintenance is important because it provides an indication that the participants can use the behavior in the absence of the conditions required to teach them. Too often in our research endeavors, when the investigation is concluded, the effects conclude as well. Researchers leave hoping the effects will be maintained, but with little evidence that this has occurred.

Assessment of social validity requires that the researcher invite participants and experts to examine the goals, procedures, and outcomes of the study. Not all studies we reviewed called this practice social validity, though we classified studies in which participants were asked specific questions about the intervention. In all, 6 (17%) studies assessed social validity. While in all of the studies using social validity the participants were asked to comment on the study, we could find no studies that referred to experts to judge the appropriateness of the goals, procedures, and outcomes of the studies.

Summary

The preceding review of studies has highlighted the contributions of behavior analysis to the pedagogical knowledge base in physical education. From the content of studies reviewed in this article, it is clear that the majority of behavioral studies
have focused on interventions designed to improve teaching behaviors. Several of these studies were the products of systematic replications. The most studied instructional intervention was the use of peers as mediators of instruction for both typically functioning students and students with disabilities. One important feature of this research and the studies in teacher preparation is that they have been conducted by a number of authors from different research programs. Such diversity of replication strengthens the confidence in the generality of results.

We have three recommendations for behavior analytic researchers and for reviewers of single-subject research. First, measures of treatment integrity should become a standard expectation for the assessment of behavioral research studies. Such measures should be accompanied by measures of reliability. In short, the same standards for measurement of the dependent variable ought to be applied to the independent variable. Second, assessment of maintenance and generality should become more commonplace in behavioral studies in physical education. The studies we reviewed show little evidence that the changes made in the presence of the independent variable have been maintained or generalized. Finally, we recommend that future research discuss the theory behind the interventions that are used in their investigations.

One limitation of this review is the number of studies that were excluded because of the inclusion criteria. In an earlier review of behavior analysis in physical education in sport, Siedentop and Taggart (1984) reported nearly 100 studies (i.e., dissertations and theses) conducted in the 1970s under the supervision of Brent Rushall and Daryl Siedentop at their respective institutions. A considerable number of studies published in the coaching literature that report on teaching sports skills were excluded from this review (Lee, 1993). For example, Garry Martin’s research program at the University of Manitoba has produced several published papers focusing on technique improvement (Martin, 1992, 1997). These studies, like those produced in Rushall’s program, probably have good generality to physical education. Another limitation is that the papers included in this review do not encompass all the possible outlets for published research of this kind. In addition, most of the journals selected for review were first published after 1970 and the absence of pre-1970 papers is clearly an artifact of our journal selection.

References

Cooper, J.O., Heron, T.E., & Heward, W.L. (1987). *Applied behavior analysis*. Columbus,
OH: Merrill.


Ratliffe, T. (1986). The influence of school principals on management time, and student


**Notes**

1 In this paper the term behavior analysis has been used to refer to the philosophy of radical behaviorism, the experimental analysis of behavior and applied behavior analysis.

2 Skinner’s (1950) paper “Are theories of learning really necessary?” made the case against the need for hypothesis testing to establish theories of cognition. His paper has often been erroneously interpreted to mean that behavior analysis is atheoretical and eschews theory. In reality, behavior analysis relies on inductive processes rather than deductive ones to develop theory.

3 Verbal behavior is the term used to describe a class of behaviors that includes talking and thinking.

4 Project Follow-Through was a large-scale, national, longitudinal evaluation study conducted in the 1970s of over 20 approaches to teaching economically disadvantaged children K–3.

5 Generalization can refer to both stimulus and response generalization. Here it is used to refer to response generalization. For a more thorough discussion of these terms, see Cooper et al. (1987).