The Effectiveness of Intervention on the Behavior of Individuals With Autism: A Meta-Analysis Using Percentage of Data Points Exceeding the Median of Baseline Phase (PEM)

Hsen-Hsing Ma

Behav Modif 2009; 33; 339 originally published online Mar 11, 2009; DOI: 10.1177/0145445509333173

The online version of this article can be found at:
http://bmo.sagepub.com/cgi/content/abstract/33/3/339

Published by:

SAGE
http://www.sagepublications.com

Additional services and information for Behavior Modification can be found at:

Email Alerts: http://bmo.sagepub.com/cgi/alerts

Subscriptions: http://bmo.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations http://bmo.sagepub.com/cgi/content/refs/33/3/339
The Effectiveness of Intervention on the Behavior of Individuals With Autism

A Meta-Analysis Using Percentage of Data Points Exceeding the Median of Baseline Phase (PEM)

Hsen-Hsing Ma
National Chengchi University, Taiwan

The aim of the present study is to demonstrate the percentage of data points exceeding the median of baseline phase (PEM) approach using data on autism treatment for illustrative purposes to compare the effectiveness of different interventions on the problem behaviors of individuals with autism. Electronic databases such as The ProQuest and Google were searched. A total of 163 articles were located, producing 1,502 effect sizes. The results demonstrate that five highly effective intervention strategies were priming, self-control, training, positive reinforcement and punishment, and presenting preferential activities. The least effective strategy was to teach perspective-taking skills. The PEM approach is recommended for use in meta-analysis for single-case experimental designs.

Keywords: autism; meta-analysis for single-case experimental designs; PEM approach

As early as in the 1970s, at least seven states in the United States had begun lawsuits against the Department of Education, for failure to provide school programs for participants with autism (Koegel & Rincover,
Such action implies the importance of intervention in support of individuals with autism in school because of their relative lack of appropriate behavior and relative abundance of inappropriate behaviors.

Several sets of criteria have been employed to diagnose children with autism, including the Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (American Psychiatric Association, 1994), the National Society for Participants with Autism (Ritvo & Freeman, 1978), or the Childhood Autism Rating Scale (Schopler, Reichler, & Renner, 1988). The typical syndromes of participants with autism are mainly: (a) qualitative impairments in social behavior (lack of awareness of the existence or feelings of others, lack of imitation of social behaviors, lack of active participation in social interactions or play); (b) qualitative impairments in verbal behavior (lack of normal development of language, echolalia, pronoun reversal, lack of eye contact, failure in making initiative or supplying feedback during conversation); and (c) stereotyped/self-stimulatory/ritualistic behaviors (meaningless repetitive movement of certain parts of the body, persistent preoccupation with parts of objects, or adherence to nonfunctional routines or rituals; American Psychiatric Association, 1994). Additionally, some individuals with autism also emit disruptive behaviors, such as aggression, noncompliance, tantrums, property destruction, and self-injury.

A number of reviews have been conducted on the treatment of autism. McLaughlin-Cheng (1998) used meta-analysis to synthesize 10 studies using between-group designs and found that participants with Asperger syndrome had a better overall performance on intelligence and cognitive measures as well as on measures of adaptive behavior abilities, such as communication, self-care skills, social appropriateness, and emotional self-regulation than those with autism. The results of Campbell’s meta-analysis (2003), which analyzed the studies using single-case experimental designs, indicated that behavior treatments (aversive, positive, combination, and extinction) were found to be significantly effective in reducing problem behavior in individuals with autism, however, no significant difference was found between the four treatments. Campbell’s classification of the treatment was broad. The present study’s aim was to classify the treatments and the dependent variables more specifically.

Mastropieri and Scruggs (1985-1986) used percentage of nonoverlapping data (PND) as a tool to calculate the effect size of experimental treatment for the meta-analysis of within-subject experimental designs. The PND is the percentage of data points in the treatment phase over the highest point of the distribution in the baseline phase (or below the lowest point of data points in the baseline phase if the undesirable behavior is expected to
decrease after the intervention is introduced). However, the PND approach has a serious problem in that if one or more data points in the baseline phase reach ceiling/floor level, then the PND scores will be 0%, although the treatment effect exists on visual inspection.

Ma (2006) has suggested the use of the PEM (percentage of data points exceeding the median of baseline phase) approach to improve this weakness of the PND approach. Ma’s study showed that the PEM had a higher Spearman correlation with original authors’ judgment than did PND. This result was also confirmed by Gao and Ma (2006) and Chen and Ma (2007). To compute the PEM scores, one needs only to draw a horizontal median line in the baseline phase. This horizontal median line will hit the median when the number of data points in the baseline phase is odd and fall between the two middle points if the number of data points is even. The median line will then stretch out horizontally to the treatment phase. Then the percentage of data points of treatment phase above the median line can be calculated. If instances of the undesired behavior are expected to decrease after the intervention is introduced, then the PEM score will be the percentage of data points below the median line in the treatment phase.

The inappropriateness of applying the methods used in the meta-analysis of between-groups researches to that of single-case experimental designs was discussed by Ma (2006). Parker and Hagan-Burke (2007) recently compared the validation of the PND, PEM, and the Improvement Risk Difference (IRD) approaches and suggested using IRD for the calculation of effect sizes for single case research, as supplements to visual analysis. However, the IRD approach is not necessarily appropriate for the calculation of the effect size of within-subject designs because of the following considerations.

1. Parker and Hagan-Burke (2007) noted: “The IRD is most commonly published as a clinical outcome measure in ‘evidence-based practice’ medical research, where it is termed ‘risk difference’ or ‘risk reduction’” (p. 921). They cited Walter’s (2000) study. However, Walter declared

   We limit attention to situations where we wish to summarize the difference between two groups with respect to a binary outcome, with event rated \( P_1 \) and \( P_2 \). Typical examples include: a clinical trial with treated and control groups, with an outcome event such as cure or death. (p. 931)

The IRD approach is appropriate for the calculation of the effect size of an experimental design, in which the distribution of the residuals is independent within group and between groups but not appropriate for the within-subject design, of which the data are dependent within phase and between phases.
2. The IRD approach is not able to allow for “efficient estimation in small samples” (Walter, 2000, p. 933), therefore inappropriate for the single-case experimental designs where the number of observation points in a phase is mostly small. The average length of treatments found in the intrasubject studies was about 10 points.

Campbell (2004) has used four methods to calculate the effect sizes of a single-case experimental design. The four methods were not considered in the present study because of the following reasons.

1. By the method of mean baseline reduction, which is calculated by subtracting the mean of treatment observations from the mean of baseline observations then dividing by the mean of baseline observations and multiplying by 100, it is difficult to have a precise reproducing the original value of data points because: (a) the graduation marked in the axis of ordinates (y-axis) is rough, usually at a regular distance of 10% extending from 0% to 100%, (b) the filled circle (size of each data point) in the figure display of a study is normally large, and (c) not all authors reported the value of mean of each phase including baseline and treatment phase. Although drafting tools have been developed to convert graphs back to raw data (Skiba, Casey, & Center, 1985-1986), the process of calculating is not as easy as that of the PEM method.

2. By the method of percentage of zero data, which is calculated by locating the first data point in the treatment phase that reaches zero and calculating the percentage of data points recorded in the treatment phase (including the first zero) that remain at zero, a lot of information will be lost, that is, the data points before the first zero data will be ignored and the figure displays containing no zero data must be excluded from the meta-analysis.

3. Although the regression-based $d$ statistic (Allison & Gorman, 1993) tried to address the issue of trend in the meta-analysis of single-case experimental designs, strictly speaking, it is a parametric statistic because it has to estimate the parameters of the effects of treatment on level, the effects of trend, and the effects of treatment on slope, and therefore not appropriate to be used for the meta-analysis of single-case experimental designs, because the data of a single-case experimental design can not satisfy the assumption of parametric statistic that the distribution of the residuals must be independent. Additionally, as the number of data points in a single-case experimental design is usually small, regression estimate based on so few data points must be suspect (Center, Skiba, & Casey, 1985-1986).

The aim of the present study is to demonstrate the PEM approach using data on autism treatment for illustrative purposes to (a) compare the effectiveness of different interventions, (b) to find out which problem behaviors are
easier to be improved and which are more difficult to be changed, and finally, (c) to discern whether study characteristics, such as mental development, gender, and age of the participants, setting of intervention, intervener, and type of experimental design, influence the effectiveness of an intervention.

Method

Data Collection of the Pilot Study

Electronic databases, such as The ProQuest Educational Journal, ProQuest Dissertation Consortium, ERIC, and Google were searched for researches investigating the effectiveness of interventions intending to improve behaviors of participants with autism. The keyword in the searches was autism. A hand search of the Journal of Applied Behavior Analysis was also conducted. Additionally, usable empirical articles were traced from the references of the located studies.

Studies that met the criterion that the data of baseline and treatment phases of a reversal or a multiple-baseline design were graphically displayed for individual participants in a time series format enabling the computation of PEM scores were included in this synthesis. Studies that used an AB design were excluded because such a design lacks internal validity, and it is not possible to rule out alternative interpretations of a result.

Altogether, 163 articles were located. The list of selected studies included in the meta-analysis is omitted in the section “reference” due to limitations of space. It is available from the author on request.

Coding of Data

The data to be coded were author(s) and publication year; categories of independent variable and dependent variable; name, age, intelligence, and gender of participant; treatment agent; setting, design, and first or second pair of baseline-treatment phase of experiment. The independent and dependent variables are grouped as follows.

Categorization of independent variables

1. **Systematic desensitization**: A strategy to reduce fears of participants with autism during common experiences. This category includes gradual exposure to fearful situations.
2. **Priming**: Helping the child prepare for upcoming activities. This category includes parent or special education staff member preview of the
classroom assignments done by the child with provision of reinforcement before they are presented in class; measuring physical appearance and personal care according to a personal appearance index; video priming for upcoming events or activities in transition.

3. **Self-control**: This package includes self-instruction, self-recording, and self-reinforcement. Self-management is a synonym of self-control.

4. **Training**: Training packages normally include four steps: Instruction, modeling/demonstration, practice/exercise, and feedback/reinforcement. This category includes categorization strategy training (sorting laundry); training gestures and verbal responses; functional communication training; training in sign language; conversation skills training; switch training (training participants with autism to use switches to activate prerecorded messages to communicate their requests); incidental teaching, enhanced milieu teaching, or embedded instruction (this kind of teaching begins with a child’s verbal or nonverbal requests, followed by a specific sequence of prompts and corrections as needed, and ends with positive feedback, expansion of the child’s utterance, and access to the child’s requested object); pivotal response training for symbolic play; naturalistic teaching coupled with use of a voice output communication aid which provides messages to be chosen for initiation or response; dry bed training (consisting of hourly awakenings, urine alarm electrode-sheet, praise for dry bed and reprimand for bed-wetting) to eliminate nocturnal bed-wetting; joint attention training.

5. **Positive reinforcement and punishment**: Positive reinforcement for desirable behaviors coupled with punishment for undesirable behaviors. This category includes praise and edible reinforcement for eye contact and punishment with functional movement (the child is required to sustain each posture of head-up, -down, or -straight for 15 seconds if the participant avoids eye-contact); rewards for preferred activities without work if the child did not ingest lethal pills and punishment with shoe polishing work if the child ingested lethal pills; using obsessions (obsessions are defined as objects or concepts with which the participants are intensely preoccupied, and continually seek out, request, talk about, or write about at home, in the after-school program of the clinic, or in school, such as plastic toy helicopters) as reinforcers for appropriate behaviors and time out as punishment.

6. **Presenting preferential reinforcers**: Presenting activities, tasks, or reinforcers that the participant with autism prefers to elicit his/her willingness and interest to engage in social interaction. This category includes interspersing the target task with varied tasks; presenting varied reinforcers in the training of academic tasks; incorporating thematic ritualistic behaviors preferred by the participant with autism into games to facilitate social play; allowing the participant to make choices through
picture exchange for preferred items during typical classroom play routines; providing choice opportunities for game play; using obsessions as reinforcers; using obsessions as tokens; incorporating echolalia in a task response (the experimenter asks the child to echo the name on the label of an object and then places two objects before the child and asks the child to hand the experimenter the labeled object); a power card strategy (synthesizing what a favorite hero/heroine would do if he/she lost or won a game and encouraging the child to generalize this strategic information across settings and events); naturalistic language teaching approach (selecting stimulus items of high interest to the child); noncontingent reinforcement (allowing the child continuous access to his/her preferred video) as a treatment for food refusal; incident teaching (when a child initiates an interaction in a natural environment, the teacher requires an elaboration of the child’s request, which is followed by the teacher’s approval, and the child is then allowed access to the requested materials, activity, or information); augmentative communication (the child is required to hand a picture over in exchange for a preferred edible reinforcer and the reinforcement is paired with a model of a natural verbal response); presenting a preferential activity schedule (demand–no activity–preferred activity (play)–and then demand); presenting entertaining music to eliminate disruptive behavior (head jerking and screaming); musical therapy consisting of listening to a song, playing rhythmic instruments, and then singing a song the words of which are composed in accordance with the guidelines for social stories.

7. **Response delay**: A few seconds of delay are permitted to encourage the child to respond.

8. **Computer-based intervention program for language training**: This category includes use of a computer-animated tutor for vocabulary and language that provides a pretest, tutorials (presenting a language lesson involving the association of pictures and spoken words) and posttest; using computer animation for enhancing communication functions (providing the children with opportunities to interact in different activities).

9. **Agent-mediated intervention**: Training parents, peers, and others on behavior modification techniques to implement treatments. This category includes peer tutoring (the tutor gives instructions or commands to the participant with autism), who is then prompted as necessary, and correct responses are reinforced with edibles and praise; peer-buddy approach (each day, the participant with autism is assigned a different buddy to stay with, play with, and talk to); the peer is instructed to initiate a play or to respond to the initiation made by the participant with autism to play; peer-incidental instruction; peer initiation and response in play with the participant with autism; rapport building (training the
caregiver to improve rapport with the child) and then to improve the child’s behaviors; peer-implemented pivotal response training consisting of modeling, role-playing, and didactic instruction; cooperative learning groups (this strategy includes grouping, assignment of team roles, group activities, and group social-skill training); training peers how to play with the participant with autism.

10. **Stimulus control**: Using stimuli to guide appropriate behaviors. This category includes a variable intermittent schedule of supervision; gradually delaying the presence of a treatment agent; providing a schedule of events or visually cued instruction; a script (consisting of statements and questions to guide the child to verbalize initiations or responses); a picture exchange communication system; visual cues (presenting a flash card with the target word to elicit self-initiated verbalizations; signaling changes in activities; scheduled awakenings).

11. **Social story**: A story consisting of four types of sentences that are (a) descriptive (defining a social setting and what people typically do in a particular situation), (b) directive (directing an individual to engage in an appropriate response in a defined situation), or represent (c) perspective (describing the perspective or reaction of others to a given situation), (d) control (providing analogies with similar actions and responses using nonhuman subjects). Such a story is written to help individuals with autism to gain insights into what others are thinking or feeling and to teach specific social skills as alternatives to problem behaviors, for example, to use quiet language to ask for help rather than crying, screaming, and the like in a frustrating situation.

12. **Punishment**: Imposing an aversive stimulus on the participant or withdrawing a positive reinforcer from the participant. This category includes sharply saying “No,” and if necessary slapping the participant briskly on the hands at the time the child begins to engage in self-stimulatory behavior; time out (letting the participant face away from the reinforcing environment); over-correction (asking the participant to exaggeratedly repeat the self-stimulatory action on the part of his or her body for a certain period); response-blocking; response cost; attention extinction; sensory extinction.

13. **Modeling**: Modeling the desirable behavior so that the participant can imitate that behavior. This category includes video modeling (presenting scripted video conversations to the child); demonstration; video self-modeling (a video is created of a role-playing movie, in the movie a participant with autism is shown socially interacting with peers in accordance with a script, then typical and positive social interactions are selected and edited for a 3-minute video and shown to the participant with autism); vivo modeling (observing another person engaging in a target behavior and subsequently imitating); peer modeling.
14. **Positive reinforcement**: This category includes reinforcing the appropriate behavior with primary reinforcers, secondary reinforcers, or preferred activities; intermittent reinforcement; gradually fading reinforcement.

15. **Differential reinforcement of others (DRO)**: Differential reinforcement of behavior other than the problem behavior. This category includes Differential Reinforcing Alternative appropriate play if the child did not engage in self-stimulation; a package consisting of DRO, response cost, and diaphragmatic breathing to relax; differential reinforcement of alternative appropriate verbal responses and extinction of perseverative verbal response.

**Categorization dependent variables**

1. **Social interaction skills**: Initiating, responding, keeping the interaction going, and verbal interactions. This category includes community skills, such as shopping; engaging in social play; cooperative play; social communication skills including appropriate eye gaze, facial expression and affect, nonverbal mannerisms, voice volume, and perseveration of topic.

2. **Language abilities**: Language abilities or behaviors including conversational speech (answering the question and providing an appropriate question). This category includes language comprehension; expressive labeling; receptive labeling; vocabulary learning; sign language (as the signs appear in a standard sign language dictionary); instruction following; expressive use of prepositions; reading skills.

3. **Attentions**: Making eye contact and paying attention to others during conversation. This category includes listening; showing enthusiastic affect during engagement in tasks; time spent on shared interest during conversation.

4. **Stereotyped behaviors**: Markedly restricted, repetitive, and stereotyped patterns of behavior, interests, and activities. This category includes self-stimulation (repetitive movements that do not appear to serve an adaptive function, such as rocking, stereotyped gazing, flapping, hand-mouthing, and object stimulation); repetitive thematic ritualistic activities with a certain object, topic, or theme.

5. **Abilities other than language ability**: Abilities or behaviors other than language ability. This category includes completion of an educational task; imitation; academic performance; appropriate play with a toy including role playing, make-believe transformation in imaginary play, and persistence during playing; on-task behavior while engaging in the assigned task; toilet use; sorting laundry; acting on-schedule (engaging in transition from one schedule to another); daily living skills; independent play; a hand-washing task; discrimination learning; self-initiations of urination; self-care skills for personal appearance; preacademic behavior.
6. **Social responses to others**: This category includes providing feedback or responding to questions; giving compliments; giving social reinforcements to others; appropriate perseveration of topic; appropriate voice volume; appropriate affective responses (displaying contextually appropriate facial, verbal, postural, and gesture responses during the interaction).

7. **Inappropriate verbal behaviors**: This category includes echolalia (meaningless repetition or echoing of verbal utterances made by a communication partner); irrelevant speech.

8. **Other inappropriate behaviors**: Undesirable behaviors other than stereotyped, repetitive behaviors. This category includes pill ingestion; self-injury; off-task; tantrums; aggression; social avoidance behavior; nonengagement; disruptive behavior including crying, screaming, hitting, falling off a chair with force; not following instructions; sleep terror (episode of sudden arousal from slow wave sleep and usually accompanied by upsetting, piercing screams, sweating, and rapid heartbeat); accidents of urinary incontinence; phobic behaviors; talk-out; ruminating/vomiting.

9. **Taking initiative**: Making initiation to begin a social interaction. This category includes asking questions; verbally or nonverbally making requests; inviting others to do something; offering or sharing with others; offering assistance; greeting others; unsolicited verbalization; initiated touch; commenting about a game or activity; securing the attention of a partner.

10. **Perspective taking**: In the theory of mind, this indicates an ability to determine mental states to explain and predict the behavior of another person.

**Designs.** A reversal design within a multiple-baseline design was treated as a reversal design. In a BAB design, the baseline phase (A) was used to calculate the median line, and the data points of the first treatment phase (B) above the median line were calculated to form the PEM scores.

**Authors’ judgments.** Authors’ conclusions of the overall effectiveness of a treatment were coded as classified, 2 = *effective or highly effective*, 1 = *moderately effective*, and 0 = *questionably or not effective*. As it was difficult to distinguish between *questionable* and *no effect*, *questionably* and *not effective* were pooled together.

**Settings.** Intervention settings were classified as home, institution (including clinic, psychiatric hospital, rehabilitation institute, residents’ living unit, teaching-family model group home, learning center, autism research
and training center, adult service program for people with developmental and behavioral disorders, and therapy room), school (including classroom, experimental room, laboratory or cafeteria in an university, and after-school behavior management program), and other places (including, community, convenience store, and museum).

**Agents.** The agents who implemented the treatments were classified into eight categories: (a) staff (including caretaker, caregiver, and supervisor), (b) author (including trainer and researcher), (c) parent (including adult promoter, mother, and teaching parents), (d) electronic teaching aids (including computer and video), (e) teacher (including instructor, special education teacher), (f) peer (including nondisabled child, sibling, and tutor), (g) therapist (including treatment provider and clinician), and (h) research assistant (including teaching assistant, experimenter, and undergraduate or graduate students).

**The intelligence of a participant.** The intelligence of a participant with autism was coded as mentally retarded or as normal.

**Ages.** Ages of participants were coded by year and month, for example, 4 years and 6 months was coded as 4.5 years. If there was no information about the ages of the individual participants but only the mean age of the group, then the mean age was used in the coding for each participant. Ages were then divided into five groups: <7 years, 7 to 12 years, 13 to 15 years, 16 to 18 years, and >18 years old.

**Calculation of Effect Size**

The effect size was only calculated when a baseline immediately before the treatment was available. In the case that a phase, no matter whether it was effective or not, was found between baseline and treatment phase, the effect size was abandoned because the effect of the treatment had been contaminated by the inserted phase. Pairs of baseline-treatment after the second pair in a reversal design were treated as a second pair. The effect sizes of the follow-up or generalization phase were not calculated except in cases where the generalization phase was included in the training and was preceded by a baseline phase, such as in Experiment 3 in Taylor and Harris’s (1995) study. Multiple treatment designs, such as a BCBC-design, where B and C denoted different treatments, were excluded from the calculation of the effect size because of the confounding of effects. However, if
the effect of a new treatment was tested, the conventional routine treatment phase was regarded as baseline phase.

Only the effect sizes of treatment on the target behaviors of the participants with autism were calculated. Behaviors that were ancillary or unrelated to the target behavior of the participants with autism as well as the target behaviors of normal peers during the interactions with the participants with autism were excluded from the calculation of the effect size. Although it is important to be able to generalize a treatment effect to real-life situations, the effect of generalization, maintenance, or follow-up phase was ignored in the present study, because of the absence of a preceding baseline phase.

Results

Two effect sizes with a mean of 0.85 in an article (Luiselli et al., 1994) were excluded from further analysis because they resulted from the treatment of ruminative vomiting by means of several procedures (dietary control, satiation, mealtime arrangement, pacing, ruminating- and vomiting-contingent consequences, weigh-ins, and medication) implemented concurrently in a multicomponent program, which were difficult to classify into a category of independent variables in the present study.

In the present study, 163 articles were located, which produced 1,502 effect sizes. The grand mean of the 1,502 effect sizes was .87 (SD = 0.25). According to the criterion of Scruggs, Mastropieri, Cook, and Escobar (1986), a PND more than 90 is considered as highly effective, 70 to 90 as moderately effective, 50 to 69 as mildly or questionable effective, and <50 as ineffective. The present study adopted this criterion and judged the grand mean of 0.87 of the 1,502 effect sizes as near highly effective treatment for improving the behaviors of participants with autism.

Normally, the effect sizes measured from a single study are not independent. To test whether the residuals of the outcome of 1,502 effect sizes were independently distributed, the lag 1 autocorrelation was calculated and found to be significant, \( r = .36, \) standard error = .03, \( t(1, 500) = 13.81. \) It indicated that the data had serial dependence, and hence, the assumption of the independent distribution of residuals was violated. Therefore, it was more suitable to apply nonparametric instead of parametric statistics to test the significance of the differences between categories.

However, if the effect sizes of each located study were averaged to form an averaged effect size to represent the effect size of that study, it was found
that the mean PEM scores of the 163 studies was 0.89 with a standard deviation of 0.15. Because the residuals of the 163 effect sizes were independently distributed (the lag 1 autocorrelation of residuals was $-0.01$, with a standard error of $0.08$, $p > .05$), a $t$ test for a sample of single group showed $t(1, 63) = 34.58$, $p < .001$. This result indicated that a mean PEM score of 0.89 for the 163 studies was significantly different from the null hypothesis of .5.

Reliability of Coding

About one third of the located studies including 497 effect sizes were randomly selected as samples for the calculation of coding reliability. Reliability was calculated by the formula: $([\text{Sum of agreement and disagreement}] - \text{number of disagreement})/\text{Sum of agreement and disagreement}$. Reliability of PEM scores $= (497 - 64)/497 = .87$. Reliability of judgment of original authors $= (497 - 123)/497 = .75$. To let the reliability of the coding approach 1.00, two assistants were then asked to code the PEM scores and the judgments of original authors for all the 1,502 effect sizes independently and the present author made the final check.

Testing the Validity of the PEM Approach

The Spearmen’s rank correlation coefficient between the judgments of the original authors and PEM scores was significant, $r(1, 500) = .40$, $p < .001$. Table 1 demonstrates that the three means of the PEM scores of

### Table 1

<table>
<thead>
<tr>
<th>Judgment of Original Author(s)</th>
<th>PEM</th>
<th>Criterion of Scruggs et al. (1986)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Not effective</td>
<td>58</td>
<td>3.9</td>
</tr>
<tr>
<td>Moderately effective</td>
<td>129</td>
<td>8.6</td>
</tr>
<tr>
<td>Highly effective</td>
<td>1,315</td>
<td>87.5</td>
</tr>
</tbody>
</table>

Note: PEM = percentage of data points exceeding the median of baseline phase.

Downloaded from http://bmo.sagepub.com at NATIONAL TAIWAN NORMAL UNIV on April 23, 2009
Table 2
Mean Effect Size of PEM Scores by Variables and Study Characteristics

<table>
<thead>
<tr>
<th>Study Characteristics</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Mr</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Independent variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Systematic desensitization</td>
<td>10</td>
<td>0.98</td>
<td>0.04</td>
<td>870</td>
</tr>
<tr>
<td>2. Priming</td>
<td>37</td>
<td>0.97</td>
<td>0.07</td>
<td>893</td>
</tr>
<tr>
<td>3. Self-control</td>
<td>45</td>
<td>0.93</td>
<td>0.16</td>
<td>801</td>
</tr>
<tr>
<td>4. Training</td>
<td>472</td>
<td>0.92</td>
<td>0.18</td>
<td>773</td>
</tr>
<tr>
<td>5. Positive reinforcement and punishment</td>
<td>55</td>
<td>0.92</td>
<td>0.18</td>
<td>854</td>
</tr>
<tr>
<td>6. Presenting preferential activities or reinforcers</td>
<td>157</td>
<td>0.91</td>
<td>0.19</td>
<td>810</td>
</tr>
<tr>
<td>7. Response delay</td>
<td>49</td>
<td>0.89</td>
<td>0.15</td>
<td>658</td>
</tr>
<tr>
<td>8. Computer-based intervention program for language training</td>
<td>45</td>
<td>0.87</td>
<td>0.19</td>
<td>686</td>
</tr>
<tr>
<td>9. Agent-mediated intervention</td>
<td>111</td>
<td>0.85</td>
<td>0.29</td>
<td>763</td>
</tr>
<tr>
<td>10. Stimulus control</td>
<td>149</td>
<td>0.84</td>
<td>0.29</td>
<td>722</td>
</tr>
<tr>
<td>11. Social story</td>
<td>49</td>
<td>0.84</td>
<td>0.23</td>
<td>688</td>
</tr>
<tr>
<td>12. Punishment</td>
<td>54</td>
<td>0.82</td>
<td>0.35</td>
<td>800</td>
</tr>
<tr>
<td>13. Modeling</td>
<td>139</td>
<td>0.79</td>
<td>0.33</td>
<td>688</td>
</tr>
<tr>
<td>14. Positive reinforcement</td>
<td>93</td>
<td>0.77</td>
<td>0.34</td>
<td>653</td>
</tr>
<tr>
<td>15. Differential reinforcement of other behaviors</td>
<td>37</td>
<td>0.73</td>
<td>0.38</td>
<td>635</td>
</tr>
<tr>
<td><strong>Dependent variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Social interactions</td>
<td>134</td>
<td>0.93</td>
<td>0.16</td>
<td>815</td>
</tr>
<tr>
<td>2. Language abilities</td>
<td>388</td>
<td>0.93</td>
<td>0.16</td>
<td>819</td>
</tr>
<tr>
<td>3. Attention</td>
<td>20</td>
<td>0.92</td>
<td>0.17</td>
<td>845</td>
</tr>
<tr>
<td>4. Stereotyped behaviors</td>
<td>75</td>
<td>0.88</td>
<td>0.26</td>
<td>803</td>
</tr>
<tr>
<td>5. Abilities other than language abilities</td>
<td>355</td>
<td>0.87</td>
<td>0.25</td>
<td>749</td>
</tr>
<tr>
<td>6. Social responses</td>
<td>95</td>
<td>0.86</td>
<td>0.27</td>
<td>661</td>
</tr>
<tr>
<td>7. Inappropriate verbal behaviors</td>
<td>28</td>
<td>0.84</td>
<td>0.22</td>
<td>586</td>
</tr>
<tr>
<td>8. Inappropriate behavior other than stereotyped behaviors</td>
<td>222</td>
<td>0.83</td>
<td>0.30</td>
<td>717</td>
</tr>
<tr>
<td>9. Making initiatives</td>
<td>168</td>
<td>0.79</td>
<td>0.32</td>
<td>679</td>
</tr>
<tr>
<td>10. Perspective-taking</td>
<td>17</td>
<td>0.67</td>
<td>0.29</td>
<td>387</td>
</tr>
<tr>
<td><strong>IQ of the participants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mentally retarded</td>
<td>984</td>
<td>0.85</td>
<td>0.27</td>
<td>538</td>
</tr>
</tbody>
</table>

(continued)
each category of judgment all fell into the range of the criterion set by Scruggs et al. (1986).
The Mean Effect Size of Independent Variables

Table 2 lists the mean, standard deviation, and mean rank of the effect sizes of the independent variables (interventions). In addition to the violation of the independence of the residuals, a Levene statistic also shows that the assumption of the homogeneity of the residuals was violated, $F(14, 1487) = 19.39, p < .001$. Throughout the present study, nonparametric statistics were used to test the significance of the difference between multiple-group means (by means of the Kruskal–Wallis one-way analysis of variance by ranks (KW ANOVA) and that of post hoc comparisons (by means of the Mann–Whitney $U$ Test).

Six categories of interventions had a large effectiveness (i.e., the mean effect size was larger than .9) on the problem behaviors of the participants with autism. They were systematic desensitization, priming, self-control, training, positive reinforcement and punishment, and presenting preferential activities or reinforcers. The remaining independent variables had a moderate effectiveness. The result of the KW ANOVA, $\chi^2(14; N = 1,502) = 37.73, p < .001$, showed that the difference between the mean rank of the effect sizes of different independent variables was significant. This result exhibited that the interventions of modeling, positive reinforcement, and differential reinforcement of other behaviors were less effective in comparison with other interventions. Post hoc comparisons of independent variables using the Mann–Whitney $U$ test had the following results: $2 > (4, 7, 8, 10, 11, 13, 14, 15); (3, 4, 5, 6) > 14; (4, 5, 6) > 7; (5, 6) > (10, 11); (4, 5) > (8, 13, 15); 6 > (13, 15)$, where the numbers are the labeling number standing before each independent variable in Table 2, where, for example, “6” stands for “presenting preferential activities or reinforcers.” The numbers within parentheses signify that there were no significant differences between the mean ranks of the effect sizes of these variables. For instance, $6 > (13, 15)$ means that the mean rank of the effect sizes of “presenting preferential activities or reinforcers” was significantly larger (more effective) than that of “modeling” and “differential reinforcement of other behaviors,” whereas the difference between that of “modeling” and “differential reinforcement of other behaviors” was not significant.

Normally, the larger the mean effect size, the larger is the mean rank of the effect size. However, sometimes there was an inconsistency between the mean rank and the mean, for example, the mean effect size of “systematic desensitization” was higher than that of “priming,” but its mean rank was lower. This phenomenon may be due to the heterogeneity of the variance of the residuals or outliers of the effect sizes.
Systematic desensitization had a large mean effect size of .98. Nevertheless, its mean rank of effect sizes was not significantly different from that of other independent variables. This may be owing to the fact that there were only 10 effect sizes of this category.

**The Mean Effect Size on the Dependent Variables**

Table 2 shows that interventions were highly effective in improving the desirable behaviors of social interactions, language abilities, and attention of the participants with autism. The mean effect sizes of these three dependent variables were all >.90. The effectiveness on the remaining variables was moderate. A Levene statistic also shows that the variances of the residuals were not homogeneous, $F(9, 1492) = 17.52, p < .001$. The result of the KW ANOVA, $\chi^2(9, N = 1,502) = 53.67, p < .001$, showed that significant difference was found between the mean rank of effect sizes of different dependent variables. Post hoc comparisons of dependent variables using the Mann–Whitney $U$ test had the following results: (1, 3) > 6; (1, 2, 3, 4, 5) > 7; (1, 2) > 8; 2 > (5, 6, 9); 4 > (6, 9); (1, 5) > 9; (1, 2, 3, 4, 5, 6, 7, 8, 9) > 10. It was relatively difficult to improve the ability of the participants in perspective taking. Table 2 also demonstrates that it was more effective to train the participants with autism to establish ability in social interaction as a whole rather than to seek to improve components of social interaction, such as taking the initiative or making social responses separately.

**The Effect of Other Moderators**

The influence of intelligence (IQ) of the participants with autism on the effectiveness of intervention. A Mann–Whitney $U$ test of the influence of IQ on the effectiveness of the interventions revealed a significant difference, $Z = -4.66, p = .001$, depicting that the interventions were more effective on the autistic participants with normal IQ ($N = 221$) than on the autistic participants with retardation ($N = 894$).

The influence of gender on the effectiveness of intervention. No significant difference was found in the influence of the gender of the participants on the effectiveness of the interventions. A Mann–Whitney $U$ test resulted in no significance, $Z = -.96, p = .34$.

The influence of length of treatment on the effectiveness of treatment. The average length of treatments was 12.78 sessions with a standard deviation
of 14.68. The Pearson correlation between the length of treatment and the PEM score was .034, \( p = .19 \), depicting that the length of time a treatment lasted did not necessarily produce a larger effect.

*The influence of type of experimental design on the effectiveness of treatment.* Treatment with a multiple baseline design had a mean effect size of .90 whereas that with a reversal design had only a mean effect size of .81. A Mann–Whitney \( U \) test revealed a significant difference, \( Z = -3.88, p < .001 \).

*The influence of pair order of reversal designs on the effectiveness of treatment.* Treatment of the first pair in a reversal design had a mean effect size of .82, whereas that of the second pair had only .79. A Mann–Whitney \( U \) test revealed no significant difference, \( Z = -.22, p = .82 \).

*The influence of agents of intervention on the effectiveness of intervention.* A Levene statistic demonstrated that the variances of the residuals was not homogeneous, \( F(7, 1446) = 9.17, p < .001 \). The result of the KW ANOVA, \( \chi^2(7, N = 1,454) = 34.12, p < .001 \), showed that significant difference was found between the mean rank of effect sizes of different agents. Post hoc comparisons of agents using the Mann–Whitney \( U \) test showed that the interventions were relatively more effective if they were implemented by the staff of institutions or the authors themselves than by other agents.

*The influence of age of participants on the effectiveness of intervention.* A Levene statistic revealed that the variances of the residuals were not homogeneous, \( F(4, 1299) = 11.04, p < .001 \). The result of the KW ANOVA, \( \chi^2(4, N = 1,304) = 9.29, p = .054 \), showed that no significant difference was found between the mean rank of effect sizes of different ages of participants, signifying that the effectiveness of intervention was not dependent on the age of the participants.

*The influence of setting of experiment on the effectiveness of intervention.* A Levene statistic revealed that the variances of the residuals were not homogeneous, \( F(3, 1292) = 9.29, p < .001 \). The result of the KW ANOVA, \( \chi^2(3, N = 1,296) = 17.37, p = .001 \), showed that setting had a significant influence on the effectiveness of an intervention. Post hoc comparisons using the Mann–Whitney \( U \) test revealed that interventions carried out in the homes of the participants and in institutions were more effective than those in other places.

*The influence of publication year of a located study on the effectiveness of intervention.* To test whether the publication year moderated effect size
of a treatment, a Pearson correlation was conducted. The correlation was not significant, $r(161) = -.03, p = .75$.

Discussion

The five highly effective (mean effect size > .9) intervention strategies were priming, self-control, training, positive reinforcement for desirable behavior plus punishment for undesirable behavior, and presenting preferential activities or reinforcers.

Dependent variables, of which the mean effect sizes were > .9 were social interactions, language abilities, and attention. The finding that it was more effective to train social interaction as a whole in individuals with autism rather than to train a single component, such as taking initiatives or making responses has practical implications. Social interaction is a continuous process and sometimes cyclic in nature with responses tending to reinforce the action of taking of the initiative in interaction. It is therefore preferable to train social interaction as a whole instead of by a single component. The fact that perspective-taking was the most difficult behavior to teach children with autism, as shown in Table 2, where the mean of the effect size was only .67, is worthy of mention here. Perspective taking is an abstract and sophisticated behavior (Charlop-Christy & Daneshvar, 2003). Children with autism have difficulty in making second-order belief attributions, for example, “person A thinks that person B thinks . . . ” (Baron-Cohen, 1989b). Tager-Flusberg’s (1992) study confirmed the experimental evidence that children with autism less than 8 years old do not understand sources of knowledge, beliefs, mental entities, pretence, and deceit. A deficiency in the acquisition of a theory of mind may account for their problems in communication and social interaction (Baron-Cohen, 1988). However, children with autism have a relatively good performance on visual perspective-taking tasks (Baron-Cohen, 1989a). Charlop-Christy and Daneshvar (2003) and LeBlanc et al. (2003) demonstrated that video modeling could be used to teach perspective-taking skills effectively to children with autism, but there has to be training in a sufficient number of exemplars, if a generalization to untrained tasks is to be expected. In Table 2, there were only 10 effect sizes for systematic desensitization and 17 effect sizes for the teaching of perspective taking, but these two intervention strategies have potential strengths in the treatment of autism. Future research should pay more attention to these areas.

The result that the difference between the mean effect size of the first pair of the baseline and that of the second pair was not significant implies
that the orthogonal effect as mentioned by Scruggs and Mastropieri (1998) was not significant and would not influence the effect size of the second treatment phase of a reversal design. The nonsignificance of the difference of different levels of gender and age shows the stability of effectiveness across age and gender of participants. The results that studies using a multiple-baseline design had a significantly larger mean effect size than those using a reversal design and that about 73% of the located studies employed a multiple-baseline design multiple-baseline design demonstrate that the multiple-baseline design is a favorable experimental design.

The PEM approaches are appropriate for the calculation of effect size of a single-case experimental design. Nevertheless, they still have methodological weaknesses, especially the problem of insensitivity to the magnitude of data points above the median (Ma, 2006). Future researches should focus on this issue.

References


**Hsen-Hsing Ma**, PhD, graduated from the School of Education, National Chengchi University, Taiwan and got his doctorate degree from the University of Düsseldorf, Germany. Since 1980, he has taught behavior modification and research methodology in education. His important published books in Chinese are *Theories and Techniques in Behavior Modification, Research Methods in Educational Science* and *Introduction to Educational Science*. Recently, he proposed the PEM approach for the meta-analysis of single case experimental researches and his articles are published in journals such as *Behavior Modification, Behavior Analysts Today*, and *Creativity Research Journal*.