An Alternative Method for Quantitative Synthesis of Single-Subject Researches

Percentage of Data Points Exceeding the Median

HSEN-HSING MA
National Chengchi University, Taiwan

The purpose of the present study is twofold: (a) to compare the validation of percentage of nonoverlapping data approach and percentage of data points exceeding the median of baseline phase (PEM) approach, and (b) to demonstrate application of the PEM approach in conducting a quantitative synthesis of single-subject research investigating the effectiveness of self-control. The results show that the PEM had higher Spearman correlation with original authors' judgment than PND did. The results of applying the PEM approach to synthesize the effect of self-control training on academic and social behavior showed that the treatment was highly or at least moderately effective.

Keywords: *PEM (percentage of data points exceeding the median of baseline phase); PND (percentage of nonoverlapping data); self-control; quantitative synthesis (meta-analysis) of single-subject research*

The purpose of the present study is twofold: (a) to compare the validation of the percentage of nonoverlapping data (PND) approach (Mastropieri & Scruggs, 1985-86) and percentage of data points exceeding the median of baseline phase (PEM) approach, and (b) to

---

AUTHOR'S NOTE: This research was supported by grants from the National Science Council, Taiwan. The assistance of part-time assistants, Miss Gao and Yu-jing, is appreciated. Correspondence concerning this article should be addressed to Hsen-hsing Ma, Department of Education, National Chengchi University, Wen-Shan District (116), Chi-nan Road, Section 2, No. 64, Taipei City, Taiwan. E-mail may be sent to gxyzwgcd@nccu.edu.tw.

BEHAVIOR MODIFICATION, Vol. 30 No. 5, September 2006 598-617
DOI: 10.1177/0145445504272974

demonstrate application of the PEM approach in conducting a quantitative synthesis of single-subject researches investigating the effectiveness of self-control in the field of applied behavior analysis.

In between-group research, many meta-analyses have been conducted to draw conclusions about the overall effectiveness of interventions (Lipsey & Wilson, 1993). But for the single-subject experimental researches, such work is just beginning. Researchers are searching for an acceptable statistical methodology to calculate the effect size of treatment of single-case experimental designs. Some researchers have proposed parametric statistics for this purpose (e.g., Center, Skiba, & Casey, 1985-86; Ferron & Sentovich, 2002; Koehler & Levin, 1998; Kromrey & Foster-Johnson, 1996; Marascuilo & Busk, 1988; Swanson & Sachse-Lee, 2000; Wampold & Worsham, 1986; White, Rusch, Kazdin, & Hartmann, 1989). These methodologies are carried over from conventional between-group research and would not necessarily be appropriate for single-subject studies. In addition to normality of distribution and homogeneity of variance of the residuals, a more important assumption of parametric statistics is that the residuals must be independently distributed (Myers, 1972). In the case of successive measurements over time in intrasubject designs, the assumption of independence of residuals is usually not met (Hersen & Barlow, 1976). The small number of data points in the phases of single-subject research would preclude the application of an autoregressive integrated moving average (ARIMA) model to the analysis of trend or level changes between baseline and treatment phases (Huitema, 1985). To correctly identify an ARIMA model in a time series, one needs at least 50 observations. A model identified with less than 20 data points would be fragile, and the number of data points in a phase of intrasubject research is normally less than 15 (Ma, 1979).

Mastropieri and Scruggs (1985-86) took a nonparametric approach to synthesize the effects of early intervention for socially withdrawn children evaluated with single-subject methodology and used the PND as the indicator of effect size. 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). The PND approach was then further applied by behavior analysts to synthesize the effect sizes of other variables (Mathur, Kavale, Quinn, Forness, & Rutherford, 1998; Scruggs, Mastropieri, Cook, & Escobar, 1986; Scruggs, Mastropieri, Forness, & Kavale, 1988).

The PND approach has the following advantages: (a) as it is a nonparametric approach, it can be free from the constraints of the assumptions of parametric statistics; (b) it is easy to calculate directly from graphic displays; and (c) it is easy to interpret qualitatively, as a PND of 90% and higher indicates highly effective, 70% to less than 90% represents moderate (or fair) effect, 50% to less than 70% indicates mild or questionable effect, whereas below 50% is considered as an ineffective treatment. This interpretation was based on previous comparisons of the PND scores by visual analysis (Scruggs et al., 1986).

However, the PND approach has crucial drawbacks.

1. If one or more data points in the baseline phase has reached the ceiling or floor level, then the PND scores will be 0%, although by visual inspection the treatment effect did exist. Yet, it is not unusual to find data points reaching the ceiling or floor level in the graphic displays of intrasubject researches (e.g., Koegel & Frea, 1993).
2. It might be expected that in the second baseline phase, the treatment effect noted in the first treatment phase would not abruptly drop to the level of the first baseline phase but become gradually extinct, and the curve in the second treatment phase would also rise gradually. There would, therefore, be an orthogonal slope change in the second pair of baseline treatment phases (Scruggs, Mastropieri, & Casto, 1987). In this case, the PND scores of the second treatment phase would be greatly underestimated.

In this regard, the PND approach would run the risk of making a Type II error (i.e., accepting the false null hypothesis).

To improve these shortcomings, the present author proposes a PEM approach. The null hypothesis of the PEM approach is that if the treatment has no effect, the data points in the treatment phase will fluctuate up and down around the middle line. The data points have 50% of chance of being above and 50% chance of being below the median of previous baseline phase.

The PEM score has a range of 0 to 1. The PEM score has the same meaning as the effect size. One can compute one PEM score from each

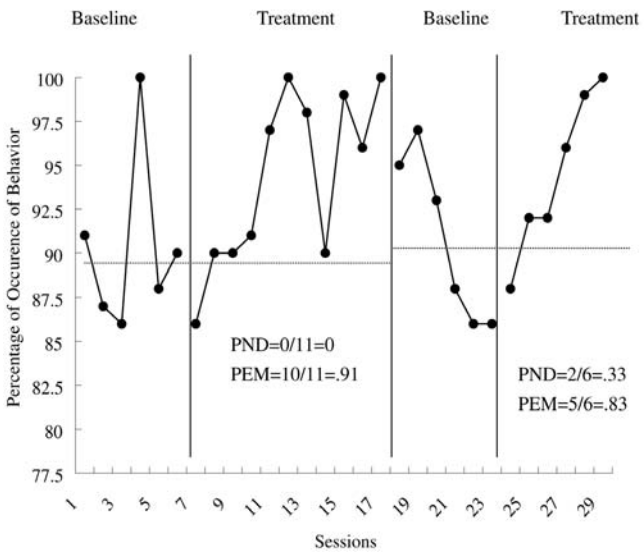

Figure 1. Demonstration of the Method of Calculating the PND and the PEM Scores and the Phenomena of Ceiling Effect and Orthogonal Slope Change

pair of baseline treatment phases. One can further calculate the overall mean effect size of each article or the mean effect size of each variable category.

In the presence of ceiling or floor data points in the baseline, as shown in Figure 1, the PEM approach is capable of computing the PEM scores that reflect the effect size whereas the PND approach cannot.

However in the presence of an orthogonal slope in the baseline treatment pair after the first treatment phase, the PEM could only have a partial improvement. Scruggs and Mastropieri (1998) have noted that this problem has rarely been encountered in the research literature. The present investigation will count the percentage of baseline treatment pairs showing orthogonal slope changes after the first treatment phase.

To demonstrate how the PEM approach can be applied in the performance of a quantitative synthesis of single-subject experimental research, research on self-control treatment were analyzed to provide an example.

There has been extensive publication of research on assessment of the effect of self-control on the undesirable behavior to be extinguished or the desirable behavior to be reinforced (Nakano, 1996). However so far, there is still no meta-analysis synthesizing the overall effectiveness of self-control investigated with single-case experimental designs.

METHOD

PROCEDURES FOR LOCATING STUDIES

The single-subject researches on self-control used in this synthesis were obtained through a computer-assisted search of the relevant databases, including EBSCOhost, ERIC, and ProQuest. Descriptors included self-control, self-instruction, self-recording, self-assessment, self-feedback, self-reinforcement, self-monitoring, and self-management. Self-instruction, self-recording, and self-reinforcement are important components of self-control. A hand search of relevant behavior analysis journals such as *Journal of Applied Behavior Analysis*, *Behavior Disorders*, *Behavior Modification*, *Behavior Assessment*, *Behavior Therapy*, *Behavior, Research and Therapy*, and *Journal of Special Education* was also conducted. Studies that meet the following criteria were included in this synthesis:

Data of baseline and treatment phases of reversal or multiple-baseline design were graphically displayed for individual participants in a time series format enabling the PND and the PEM scores to be computed. The first pair of baseline-treatment phases or the pair after that was also coded. Generalization or follow-up phase as well as treatment phase without immediate preceding baseline phase was not included in the analysis because their effect might be contaminated by the preceding phase.

PROCEDURE FOR CODING THE STUDY

Study characteristics. Variables in each of the following areas were coded:

1. Authors' conclusion of overall effectiveness of treatment (2 = *effective*, 1 = *moderately effective*, or 0 = *questionably or not effective*). Because it is hard to distinguish between questionable and no effect, they were pooled together.
2. Categorization of independent variables. Independent variables were divided into four categories: (a) self-instruction (self-statement and reading aloud the instruction are attributed to this category), (b) monitoring (synonymous terms are self-evaluation, self-recording, self-assessment, and self-checking), (c) self-reinforcement, and (d) self-control package (i.e., a composition of two or more of the above elements; synonymous terms are self-management and self-regulation).
3. Categorization of dependent variables. Target behaviors were classified into four categories: (a) academic behaviors measured as accuracy (or proficiency, grades, correct responses), (b) academic behaviors measured as task completed, (c) socially desirable behaviors (on task, appropriate behaviors, attending, desirable peer interactions, communication skills), and (d) socially undesirable behavior (aggressive behavior, disruptive behaviors, drug abuse, inappropriate communicative behaviors, off task, self-stimulations, left too early, absence, coming too late).
4. Settings. Intervention settings were classified as home, institution (including clinic and various therapeutic centers), school, and other places (including company, community, and swimming pool).
5. Interveners. They were categorized into experimenter (including treatment provider, trainer, research assistant, and instructor), staff (including therapist, facilitator, teaching parent, counselor, and clinician), teacher (including swimming coach), and tutor (including peer teacher and home tutor).
6. Participant Classifications. Participants in the present study were classified as attention deficit hyperactivity disorder, autism, brain injury, chronic alcoholic, emotional disturbance, learning disability, mental retardation, and normal (including participants with normal IQ but having behavior problems, such as disruptive, predelinquent, and socially isolated, underachieving, or having other behavior disorders).
7. Participant Sex and Age. Age was divided into five groups: (a) younger than 7 years, (b) 7 to 12 years, (c) 13 to 15 years, (d) 16 to 18 years, and (e) older than 18 years.

COMPUTATION OF TREATMENT OUTCOMES

Treatment outcomes were calculated by computing the PND scores and the PEM scores of each pair of baseline treatment phases.

Reliability. A student in a doctoral program of education serving as a part-time research assistant conducted the variable coding and calculation of the PND as well as the PEM scores. The present author checked her work, and the percentage of agreement was counted. The percentage of agreement is calculated by the formula: the number of agree divided by the number of agree plus the number of disagree. Disagreements were resolved by discussion until a consensus was reached and recalculated.

Calculation of the PEM. To compute the PEM scores, one needs only to draw a horizontal middle line in the baseline phase. This horizontal middle line will hit the median when the number of data points in the baseline phase is odd and go between the two middle points if the number of data points is even. This middle line will stretch out horizontally to the treatment phase. Then the percentage of data points of treatment phase above the middle line may be calculated. If the undesired behavior is expected to decrease after the treatment is introduced, then the PEM score will be the percentage of data points below the middle line in the treatment phase.

Figure 1 demonstrates the method of calculating the PEM and PND. There are ten points in the first treatment phase over the median line. Therefore, the PEM is 10/11 = 90.9%. And the PND = 0/11 = 0%. For an orthogonal slop change, the PEM has also a higher score than the PND.

RESULTS

From the total of 61 articles used for quantitative synthesis in the present study, 16 articles (202 pairs of baseline treatment phases) were sampled for the calculation of coding reliability. Percentage of agreement between the present author's coding and that of the research assistant was 83.65% for the coding of original authors' judgments, and 95.85% for the PND, and 94.55% for the PEM. Most of the inconsistency in coding the original authors' judgments on treatment effects was found in the category of *moderate effect*, which was coded as 1,

TABLE 1
Intercorrelations Between Original Authors' Judgment, the PND Scores, and the PEM Scores

Scores	*1*	*2*	*3*
With Pair as Unit			
1. Judgment (n = 647)		.49***	.57***
2. PND (n = 659)			.64***
3. PEM (n = 659)			
With Article as Unit (k = 61)			
1. Judgment		.47***	.59***
2. PND			.69***
3. PEM			

NOTE: The coefficients between judgment and the PND or with the PEM are Spearmen's rank correlation, but the coefficients between the PND and the PEM are Pearson's product-moment correlation because the judgment scores are coded with ordinal scale.
***p < .001

whereas *noticeable effect* (coded as 2) and *little effect or no improvement* (coded as 0) showed little confusion.

The following terms were assigned as *moderately effective*: increased but slow, variable but increasing trend, positive change but inconsistently, small increase, slightly above baseline, increased but overlap with baseline, slight increase, and increased but did not quite reach the norm.

As the coding numbers of the judgments of original authors on the treatment effects were of ordinal scale, the Spearman correlation was used to decide which approach, the PND or the PEM, had a higher consonance with original researchers' judgment on treatment effect. The matrix of intercorrelation coefficients between the judgments of original researchers, the PND, and the PEM is presented in Table 1 with number of effect sizes in parentheses.

Table 1 shows that the PEM has a higher correlation with the original authors' judgment than that of the PND with original authors' judgment, no matter whether it is calculated with the sample of pairs of baseline treatment phase or with sample of articles having only one average value of effect size.

The mean of 659 PEM scores is .87 with a standard deviation of .24. The mean of 659 PND scores is .61 with a standard deviation of .39.

To respond to the critics that effect sizes in an article are not independent, the effect sizes of each article are averaged to form a single average effect size. It was found that the mean of 61 PEM scores is .90 with a standard deviation of .13. The mean of 61 PND scores is .67 with a standard deviation of .26.

In searching the change in the orthogonal slope after the first treatment phase, only two obvious orthogonal slope changes in the second baseline phase out of 61 articles had been found. They were found in the diagrams for Participants 1, 2, and 7 in Figure 1 of Olympia, Sheridan, Jenson and Andrews (1994), and Student 4 in Figure 1 of Koegel and Koegel (1990). To investigate whether the orthogonal slope change threaded the effect size of the second baseline treatment pair, the averaged effect size of the second baseline treatment pair from 82 ABAB designs was subtracted from that of the first baseline treatment pair. The results showed that the difference of two baseline treatment pairs was 0.033 for the original authors' judgment, 0.013 for the PND, and 0.009 for the PEM. The results indicate that the effect size of the second baseline treatment pair is higher than that of the first one, and the differences were small; therefore, the effect of orthogonal slope change can be negligible.

More specific breakdown of the effect of self-control by the PEM, the PND, and original authors' judgments are given in Table 2.

Under the condition of unequal size, the heterogeneity of variance would cause serious consequences (Scheffe, 1961), and it can be seen in Table 2 that the sample sizes of subcategories are not equal. Accordingly, score differences from various study characteristics could not be compared by means of parametric statistics.

The *n* in Table 2 designates the number of baseline treatment pairs as the unit of analysis with the exception of overall effect with article as unit. Table 2 shows that means of the PEM scores of each subcategory of variable range from .80 to .98. When the criterion of Scruggs et al. (1986) was used, all of the subcategory of variable would have at least moderate (between .70 and .89) or high effectiveness (higher than .9).

Scruggs et al. (1986) applied nonparametric tests (Kruskal-Wallis one-way analysis of variance by ranks) to compare the PND score dif-

TABLE 2
Effect Size by Study Characteristics

	PEM			*PND*			*Authors' Judgment*		
Variable	M	SD	N	M	SD	N	M	SD	N
Overall effect									
With baseline-treatment pair as unit	0.87	0.23	659	0.61	0.39	659	1.67	0.65	651
With article as unit	0.9	0.13	61	0.67	0.26	61	1.8	0.42	61
t test	$t(60) = 24.68^{***}$			$t(60) = 19.85^{***}$			$t(60) = 33.94^{***}$		
Intervention									
Self-control package	0.82	0.26	258	0.51	0.41	258	1.57	0.67	251
Self-instruction	0.88	0.23	91	0.77	0.34	91	1.77	0.62	91
Self-monitoring	0.9	0.21	301	0.64	0.37	301	1.73	0.64	296
Self-reinforcement	0.94	0.17	9	0.94	0.17	9	2	0	9
Kruskal-Wallis ANOVA by ranks	$\chi^2 (3, n = 659) = 17.83^{***}$			$\chi^2 (3, n = 659) = 38.33^{***}$			$\chi^2 (3, n = 647) = 22.22^{***}$		
Behavior (dependent variable)									
Academic behavior (accuracy)	0.89	0.22	221	0.68	0.39	221	1.71	0.56	216
Academic behavior (work completed)	0.8	0.3	77	0.49	0.36	77	1.51	0.84	77
Social behavior (desirable)	0.88	0.22	266	0.6	0.39	266	1.68	0.67	266
Social behavior (undesirable behavior reduced)	0.84	0.25	95	0.55	0.42	95	1.76	0.61	88
Kruskal-Wallis ANOVA by ranks	$\chi^2 (3, n = 659) = 17.95^{***}$			$\chi^2 (3, n = 659) = 17.41^{***}$			$\chi^2 (3, n = 647) = 4.80^{ns}$		
Setting									
Home	0.98	0.06	33	0.91	0.21	33	2	0	33
Institution	0.82	0.27	147	0.49	0.4	147	1.56	0.79	147
School	0.88	0.23	416	0.64	0.39	416	1.65	0.65	404
Other places	0.84	0.22	51	0.48	0.37	51	1.98	0.14	51
Kruskal-Wallis ANOVA by ranks	$\chi^2 (3, n = 647) = 25.93^{***}$			$\chi^2 (3, n = 647) = 34.18^{***}$			$\chi^2 (3, n = 635) = 27.29^{***}$		
Subject age									
Younger than 7 years	0.91	0.20	15	0.54	0.44	15	1.6	0.83	15
7-12 years	0.86	0.24	367	0.59	0.38	367	1.56	0.71	362
13-15 years	0.88	0.26	104	0.62	0.42	104	1.87	0.47	97
16-18 years	0.89	0.23	32	0.58	0.46	32	2	0	32
More than 18 years	0.88	0.21	123	0.65	0.4	123	1.77	0.6	123
Kruskal-Wallis ANOVA by ranks	$\chi^2 (3, n = 647) = 16.28^{***}$			$c^2 (3, n = 647) = 4.41^{ns}$			$c^2 (3, n = 635) = 37.82^{***}$		
Subject sex									
Female	0.88	0.22	190	0.63	0.4	190	1.7	0.68	187
Male	0.88	0.23	323	0.6	0.39	323	1.71	0.66	321
Kruskal-Wallis ANOVA by ranks	$\chi^2 (1, n = 641) = 0.15^{ns}$			$\chi^2 (1, n = 641) = 0.94^{ns}$			$\chi^2 (1, n = 629) = 0.06^{ns}$		

(continued)

TABLE 2 (continued)

Variable	*PEM*			*PND*			*Authors' Judgment*		
	M	SD	N	M	SD	N	M	SD	N
Subject classifications									
Attention deficit hyperactivity disorder	0.93	0.08	16	0.66	0.35	16	1.81	0.4	16
Autism	0.93	0.14	37	0.57	0.44	37	1.86	0.35	37
Brain injury	0.96	0.12	16	0.94	0.14	16	2	0	16
Chronic alcoholics	0.88	0.25	4	0.88	0.25	4	2	0	4
Emotional disturbance	0.89	0.26	66	0.68	0.42	66	1.83	0.48	66
Learning disability	0.88	0.22	152	0.59	0.38	152	1.54	0.8	147
Mental retardation	0.83	0.28	128	0.65	0.39	128	1.69	0.72	128
Normal	0.86	0.23	238	0.55	0.4	238	1.65	0.61	231
Kruskal-Wallis ANOVA by ranks	χ^2 (7, n = 657) = 20.38**			χ^2 (7, n = 657) = 28.12***			χ^2 (7, n = 645) = 18.38**		
Intervener									
Researcher	0.83	0.25	126	0.5	0.41	126	1.48	0.67	126
Experimenter	0.91	0.2	127	0.73	0.37	127	1.87	0.46	127
Staff	0.82	0.27	100	0.51	0.38	100	1.68	0.69	100
Teacher	0.87	0.24	264	0.58	0.39	264	1.63	0.72	252
Tutor	0.99	0.04	28	0.97	0.1	28	2	0	28
Kruskal-Wallis ANOVA by ranks	χ^2 (4, n = 645) = 36.35***			χ^2 (4, n = 645) = 65.53***			χ^2 (4, n = 633) = 45.22***		

**p < 01. ***p < .001. *ns* = not significant.

ferences by various study characteristics. So it is reasonable to do so for the PEM. Breakdowns are given in Table 2.

The overall averaged effect size for 659 pairs of baseline treatment phases was 0.87. It was moderately effective according to Scruggs' et al. (1986) criterion. The averaged effect size for 61 independent studies was 0.9. It was significantly different from 0.5 (t(60) = 24.68, p < .0001).

The differences of four interventions were statistically meaningful, and single element of self-control package, especially the self-reinforcement, were more effective than the whole package.

For the target behaviors, self-control treatments were more successful on the academic behaviors associating with accuracy of work and on establishing desirable social behaviors than reducing undesirable social behavior and promoting completion of academic tasks.

Home and school were associated with significantly stronger outcome than institutions and other settings.

Tutors and experimenters were more effective than researchers and teachers. Overall differences between eight groups of participants appeared to be statistically significant. Participants with brain injury, attention deficit hyperactivity disorder, and autism revealed more improvements than those with emotional disturbance, learning disability, normal, and mental retardation.

Differences in the PEM scores by age level were statistically significant but not systematic. Participants younger than 7 years old showed most effective change by the self-control treatments, but participants between 7 and 12 years old showed the least.

Differences in the PEM scores by sex of participant were not statistically significant.

Table 3 displays the *mean*, *standard deviation*, and *number of pairs* of baseline treatment phase of three subcategories of qualitative judgment over the effectiveness of treatment. There were 501 pairs of baseline treatment phases, which the original authors judged as highly effective. The mean of the PEM scores of these pairs is .94, although the mean of the PND scores is .72. It seems that the criterion set by Scruggs et al. (1986) is more suitable for the PEM scores and too stringent for the PND scores.

DISCUSSION

The present meta-analysis with the PEM approach found that self-control training, either in the form of a self-control package or in the form of single element of self-control, such as self-instruction, self-monitoring, or self-reinforcement, had an effect on all four categories of behaviors. There is, up to now, no other meta-analysis of single-subject studies about self-control. The present author could only compare the findings with the results of meta-analysis of between-group studies. The results are consistent with the results of Baker, Swisher, Nadenichek, and Popowicz (1984) as well as Stage and Quiroz (1997).

TABLE 3
Comparisons of Means of the PEM and the PND Scores With Criteria Suggested by Scruggs, Mastropieri, Cook, & Escobar, (1986) at Each Level of Effectiveness Judged by Original Authors

		PEM		*PND*		
Original Authors' Judgment	N	M	SD	M	SD	*The Criterion of Scruggs et al. (1986)*
Highly effective	501	.94	.14	.72	.34	≥ .9
Moderately effective	80	.76	.24	.38	.35	≥.7 < .9
Questionable or not effective	66	.48	.33	.08	.14	< .7

The display in Table 1 indicates that the PEM scores have a higher correlation with the original authors' judgments than the PND scores do. This finding was supported by a recent study (Gao, 2004). However, only about 25% of the variance in treatment effectiveness is shared by the PND, the PEM, and authors' judgments. A possible reason that correlation was not high enough might be that there are only three scores for the authors' judgments in ordinal scale, and the PND and the PEM scores are of interval scale; therefore, the rank correlation between them could not be expected to be high. Other factors contributing to the modest correlation have to be investigated in future research. The present author thought that the ones who know best the effectiveness of treatment are the original authors of each study. Mastropieri and Scruggs (1985-86) had adopted six considerations from Parsonson and Baer (1978) to determine the effectiveness of outcome. Their six considerations contain nevertheless ambiguous terms, such as *questionable*, *adequacy*, *amount*, and *inappropriate*, and lack precise quantitative criterion for coding. Therefore, the original authors' judgments were used instead of the six considerations.

Furthermore, the PEM is free from the fatal influence of the data point, which has reached the ceiling or floor level in the baseline phase. This has been a source for concern in the use of the PND approach. Because a single extreme outlier can produce a detrimental effect on the PND score (Scruggs et al., 1987). Researches with results that have a data point reaching ceiling or floor in the baseline phase were found in Billings and Wasik (1985); Blick and Test (1987);

Brigham, Hopper, Hill, Armas, and Newsom (1985); Burgio, Whitman, and Johnson (1980); Burgio, Whitman, and Reid (1983); Carr and Punzo (1993); Dunlap and Dunlap (1989); Glomb and West (1990); Gumpel and David (2000); Kern, Ringdah, Hilt, and Sterling-Turner (2001); Kern-Dunlap et al. (1992); Kissel, Whitman, and Reid (1983); Koegel, Keogel, Hurley, and Frea (1992); Koegel and Frea (1993); Martin and Manno (1995); McKenzie and Rushall (1974); Olympia et al. (1994); Stahmer and Schreibman (1992); Swanson (1981); Wilson, Leaf, and Nathan (1975); and Wood, Murdock, and Cronin, (2002).

There are some methodological problems not addressed by the PEM approach

1. Insensitivity to the magnitude of data points above the median. Scores of 100% could be obtained whether all treatment scores were just slightly or substantially higher than the median of the preceding baseline.
2. Trend and variability in data points of the treatment phase were not considered. This problem can only be alleviated if researchers terminate treatment phase after the observations are stable.
3. The problem of applying a *t* test to examine the significance of overall mean of the PEM scores. The prerequisites for applying a *t* test, which belongs to parametric statistics, are normality, homogeneity of variances, independence of the residuals, and data must be at least of interval scale. The PEM scores might not always be distributed normally; however, violation of normality would not cause serious consequences (Lindquist, 1956). The homogeneity of variances of the residuals would not be violated if two compared groups have equal size (Scheffe, 1961). In the case of the PEM, the mean of the PEM scores is to be compared with .5 of the null hypothesis, each PEM score will have a counterpart of .5, so the two groups of data have equal size. A PEM score (a percentage) is of interval scale; because the distance between units is equal, it has absolute zero point, but there is no ratio relation between units (e.g., a PEM score of 1 is not two times more effective than that of .5). To test whether the residuals of effect size of 659 pairs and 61 articles were independently distributed, the time-series analysis (ARIMA) was applied. After the time series was centered by subtraction of its mean, all first 24 lags of autocorrelation of residuals of the PEM scores of 659 pairs were significant, although none of autocorrelation of the residuals of the averaged PEM scores of 61 articles were significant. It was also the case for the PND and authors' judgments. It signified that the residuals of effect sizes in an article

would not independently distributed, but if the effect sizes were averaged to form a mean effect size for that article, then the residuals of the averaged effect sizes would be distributed independently, and it is eligible to apply the *t* test.

$$t = \frac{ES - 5}{\frac{SD}{\sqrt{N}}} \quad (1)$$

Where *ES* is the mean effect size of all articles, with each article having only one averaged effect size (independent effect size), *SD* is the standard deviation of all independent effect sizes; *n* is the number of independent effect size.

As the *t* test with Formula 1 was applied, the *t*-value ($t(60) = 24.68$, $p > .0001$) seems too large. However, in the present study, there were 501 pairs (77%) of baseline treatment phases judged by original authors as effective, 80 pairs (12%) judged as moderately effective, and 66 pairs (10%) judged as ineffective. Both results seem not to be controversial.

The present author wants to leave open for discussion the suitability of applying the *t* test to determine the significance of the overall effect of treatment. It is hoped that under empirical data-based discussions, a universally acceptable approach of quantitative synthesis of single-subject studies, which can simultaneously address these three problems, will be elicited, and then the results of empirical research of single-subject studies can be consolidated into the body of knowledge in applied behavior science.

REFERENCES

References marked with an asterisk indicate studies included in the meta-analysis.

Baker, S. B., Swisher, J. D., Nadenichek, P. E., & Popowicz, C. L. (1984). Measured effects of primary prevention strategies. *Personnel and Guidance Journal*, *62*, 459-464.

Billings, D. C., & Wasik, B. H. (1985). Self-instructional training with preschoolers: An attempt to replicate. *Journal of Applied Behavior Analysis*, *18*, 61-67.

Blick, D. W., & Test, D. W. (1987). Effects of self-recording on high school students' on task behavior. *Learning Disability Quarterly*, *10*, 203-213.

*Bornstein, P. H., & Quevillon, R. P. (1976). The self-instructional package on overactive preschool boys. *Journal of Applied Behavior Analysis*, *9*, 179-188.

Brigham, T. A., Hopper, C., Hill, B., Armas, A. D., & Newsom, P. (1985). A self-management program for disruptive adolescents in the school: A clinical replication analysis. *Behavior Therapy*, *16*, 99-115.

*Broden, M., Hall, R. V., & Mitts, B. (1971). The effect of self-recording on the classroom behavior of two eighth grade students. *Journal of Applied Behavior Analysis*, *4*, 191-199.

Burgio, L. D., Whitman, T. L., & Johnson, M. R. (1980). A self-instructional package for increasing attending behavior in educable mentally retarded children. *Journal of Applied Behavior Analysis*, *13*, 443-459.

Burgio, L. D., Whitman, T. L., & Reid, D. H. (1983). A participative management approach for improving direct-care staff performance in an institutional setting. *Journal of Applied Behavior Analysis*, *16*, 37-53.

Carr, S. C., & Punzo, R. P. (1993). The effects of self-monitoring of academic accuracy and productivity on the performance of students with behavioral disorders. *Behavioral Disorders*, *18*, 241-250.

Center, B. A., Skiba, R. J., Casey, A. (1985-86). Methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, *19*, 387-400.

*Chou, T. J., & Lin, Y. H. (1996). The effects of self-instructional training on ADHD children. *Journal of Special Education*, *11*, 239-284.

*Christian, L. (1997). Using self-management procedures to improve the productivity of adults with developmental disabilities in a competitive employment setting. *Journal of Applied Behavior Analysis*, *30*, 169-172.

*Christie, D. J., Hiss, M., & Lozanoff, B. (1984). Modification of inattentive classroom behavior. *Behavior Modification*, *8*, 391-406.

*Connis, R. T. (1979). The effects of sequential pictorial cues, self-recording, and praise on the job task sequencing of retarded adults. *Journal of Applied Behavior Analysis*, *12*, 355-361.

Dunlap, L. K., & Dunlap. G. (1989). A self-monitoring package for teaching subtraction with regrouping to students with learning disabilities. *Journal of Applied Behavior Analysis*, *22*, 309-314.

*Feldman, M. A., Ducharme, J. M., & Case, L. (1999). Using self-instruction pictorial manuals to teach child-care skills to mothers with intellectual disabilities. *Behavior Modification*, *23*, 480-497.

Ferron, J., & Sentovich, C. (2002), Statistical power of randomization tests used with multiple-baseline designs. *The Journal of Experimental Education*, *70*, 165-178.

*Foxx, R. M., & Rubinoff, A. (1979). Behavioral treatment of caffeinism: Reducing excessive coffee drinking. *Journal of Applied Behavior Analysis*, *12*, 335-344.

*Frea, W. D., & Hughes. C. (1997). Functional analysis and treatment of social-communicative behavior of adolescents with developmental disabilities. *Journal of Applied Behavior Analysis*, *30*, 701-704.

*Gajar, A., Schloss, P. J., Schloss, C. N., & Thompson, C. K. (1984). Effects of feedback and self-monitoring on head trauma youths' conversation skills. *Journal of Applied Behavior Analysis*, *17*, 353-358.

Gao, Y. J. (2004). *A quantitative synthesis of single-subject researches into the effect of behavioral modification on academic behaviors*. Unpublished doctoral dissertation, National Cheng-Chi University, Republic of China.

Glomb, N., & West, R. P. (1990). Teaching behaviorally disordered adolescents to use self-management skills for improving the completeness, accuracy, and neatness of creative writing homework assignments. *Behavioral Disorders*, *15*, 233-242.

*Glynn, E. L., & Thomas, J. D. (1974). Effect of cueing on self-control of classroom behavior. *Journal of Applied Behavior Analysis*, *7*, 299-306.

*Gumpel, T. P., & David, S. (2000). Exploring the efficacy of self-regulatory training as a possible alternative to social skills training. *Behavioral Disorders*, *23(5)*, 131-141.

*Hallahan, D. P., Marshall, K. J., & Lloyd, J. W. (1981). Self-recording during group instruction: Effects on attention to task. *Learning Disability Quarterly*, *4*, 407-413.

*Hallahan, D. P., Lloyd, J. W., Kneedler, R. D., & Marshall, K. J. (1982). A comparison of the effects of self-versus teacher-assessment of on-task behavior. *Behavior Therapy*, *13*, 715-723.

*Harris, K. R. (1986). Self-monitoring of attentional behavior versus self-monitoring of productivity: Effects on on-task behavior and academic response rate among learning disabled children. *Journal of Applied Behavior analysis*, *19*, 417-423.

*Harris, K. R., & Graham, S. (1985). Improving learning disabled students' composition skills: Self-control strategy training. *Learning Disability Quarterly*, *8*, 27-36.

*Harris, K. R., Graham, S., Reid, R., McElroy, K., & Hamby, R. S. (1994). Self-monitoring of attention versus self-monitoring of performance: Replication and cross-task comparison studies. *Learning Disability Quarterly*, *17*, 121-139.

Hersen, M., & Barlow, D. (1976). *Single case experimental designs*. New York: Pergamon.

*Hughes, C., Harmer, M. L., & Killian, D. J. (1995). The effects of multiple-exemplar self-instructional training on high school students' generalized conversational interactions. *Journal of Applied Behavior Analysis*, *28*, 201-218.

*Hughes, C., & Rusch, F. R. (1989). Teaching supported employees with severe mental retardation to solve problems. *Journal of Applied Behavior Analysis*, *22*, 365-372.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, *7*, 107-118.

*Jones, R. T., Kazdin, A. E., & Haney, J. I. (1981). Social validation and training of emergency fire safety skills for potential injury prevention and life saving. *Journal of Applied Behavior Analysis*, *14*, 249-260.

Kern, L., Ringdah, J. E., Hilt, A., & Sterling-Turner, H. E. (2001). Linking self-management procedures to functional analysis results. *Behavior Disorders*, *26*, 214-226.

Kern-Dunlap, L., Dunlap, G., Clarke, S., Childs, K. E., White. R. L., & Stewart, M. P. (1992). Effects of a videotape feedback package on the peer interactions of children with serious behavioral and emotional challenges. *Journal of Applied Behavior Analysis*, *25*, 355-364.

Kissel, R. C., Whitman, T. L., & Reid, D. H. (1983). An institutional staff training and self-management program for developing multiple self-care skills in severely profoundly retarded individuals. *Journal of Applied Behavior Analysis*, *16*, 395-415.

*Knapczyk, D. R., & Livingston, G. (1973). Self-recording and student teacher supervision: Variables within a token economy structure. *Journal of Applied Behavior Analysis*, *6*, 481-486.

Koegel, L. K., Koegel, R. L., Hurley, C., & Frea, W. D. (1992). Improving social skills and disruptive behavior in children with autism through self-management. *Journal of Applied Behavior Analysis*, *25*, 341-353.

Koegel, R. L., & Frea, W. D. (1993). Treatment of social behavior in autism through the modification of pivotal social skills. *Journal of Applied Behavior Analysis*, *26*, 369-377.

Koegel, R. L., & Koegel, L. K. (1990). Extended reductions in stereotypic behavior of students with autism through a self-management treatment package. *Journal of Applied Behavior Analysis*, *23*, 119-127.

Koehler, M., & Levin, J. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods*, *3*, 206-217.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education*, *65*, 73-93.

*Levendoski, L. S., & Cartledge, G. (2000). Self-monitoring for elementary school children with serious emotional disturbances: Classroom applications for increased academic responding. *Behavioral Disorders*, *25*, 211-224.

*Likins, M., Salzberg, C. L., Stowitschek, J. J., Kraft, B. L., & Curl, R. (1989). Co-worker implemented job training: The use of coincidental training and quality-control checking on the food preparation skills of trainees with mental retardation. *Journal of Applied Behavior Analysis*, *22*, 381-393.

Lindquist, E. F. (1956). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist*, *48*, 1181-1209.

*Lloyd, J. W., Hallahan, D. P., Kosiewicz, M. M., & Kneedler, R. D. (1982). Reactive effects of self-assessment and self-recording on attention to task and academic productivity. *Learning Disability Quarterly*, *5*, 216-226.

*Lloyd, J. W., Bateman, D. F., Landrum, T. J., & Hallahan, D. P. (1989). Self-recording of attention versus productivity. *Journal of Applied Behavior Analysis*, *22*, 315-323.

Ma, H.-H. (1979). *Der experimentelle Einzelfalluntersuchung in der erziehungswissenschaftlichen Forschung* [The experimental single subject research in education], Unpublished doctoral dissertation. University of Duesseldorf, Germany.

*Maag, J. W., & Peid, R. (1993). Differential effects of self-monitoring attention, accuracy, and productivity. *Journal of Applied Behavior Analysis*, *26*, 329-344.

*Marascuilo, L., & Busk, P. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, *10*, 1-28.

Martin, K. F., & Manno, C. (1995). Use of a check-off system to improve middle school students' story compositions. *Journal of Learning Disabilities*, *28*, 139-149.

Mastropieri, M. A., & Scruggs, T. E. (1985-86). Early intervention for socially withdrawn children. *The Journal of Special Education*, *19*, 429-441.

Mathur, S. R., Kavale, K. A., Ouinn, M. M., Forness, S. R., & Rutherford, R. B. (1998). Social skills interventions with students with emotional and behavioral problems: A quantitative synthesis of single-subject research. *Behavioral Disorders*, *23*, 193-201.

*McKenzie, T. L., & Rushall, B. S. (1974). Effects of self-recording on attendance and performance in a competitive swimming training environment. *Journal of Applied Behavior Analysis*, *7*, 199-206.

*Miller, M., Miller, S. R., Wheeler, J., & Selinger, J. (1989). Can a single-classroom treatment approach change academic performance and behavioral characteristics in severely behaviorally disordered adolescents: An experimental inquiry. *Behavioral Disorders*, *14*, 215-225.

Myers, J. L. (1972). *Fundamental of experimental design.* Boston: Allyn & Bacon.

Nakano, K. (1996). Application of self-control procedures to modifying type A behavior. *Psychological Record*, *46*, 595-607.

*Ninness, H. A., Fuerst, J., & Rutherford, R.D. (1991). The effect of self-management training package on the transfer of aggression control procedures in the absence of supervision. *Behavior Modification*, *19*, 464-490.

*Ninness, H. A. C., Fuerst, J., & Rutherford, R. D. (1991). Effects of self-management training and reinforcement on the transfer of improved conduct in the absence of supervision. *Journal of Applied Behavior Analysis*, *24*, 499-508.

Olympia, D. E., Sheridan, S. M., Jenson, W. R., & Andrews, D. (1994). Using student-managed interventions to increase homework completion and accuracy. *Journal of Applied Behavior Analysis*, *27*, 85-99.

*O'Reilly, M. F., Green, G., & Braunling-McMorrow, D. (1990). Self-administered written prompts to teach home accident prevention skills to adults with brain injuries. *Journal of Applied Behavior Analysis*, *23*, 431-446.

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. R. Dratochwill (Ed), *Single-subject research: Strategies for evaluating change* (pp. 101-102). New York: Academic Press.

*Prater, M. A., Joy, R., Chilman, B., Temple, J., & Miller, S. R. (1991). Self-monitoring of on-task behavior by adolescents with learning disabilities. *Learning Disability Quarterly*, *14*, 164-177.

*Roberts, R. N., Nelson, R. O., & Olson, T. W. (1987). Self-instruction: An analysis of the differential effects of instruction and reinforcement. *Journal of Applied Behavior Analysis*, *20*, 235-242.

*Rooney, K., Polloway, E. A., & Hallahan, D. P. (1985). The use of self-monitoring procedures with low IQ learning disabled students. *Journal of Learning Disabilities*, *18*, 384-389.

*Rooney, K. J., Hallahan, D. P., & Lloyd, J. W. (1984). Self-recording of attention by learning disabled students in the regular classroom. *Journal of Learning Disabilities*, *17*, 360-364.

Scheffe, H. A. (1961). *The analysis of variance.* New York: John Wiley & Sons.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research. *Behavior Modification*, *22*, 221-243.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8(2)*, 24-33.

Scruggs, T. E., Mastropieri, M. A., Cook, S. B., & Escobar, C. (1986). Early interventions for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders*, *11*, 260-271.

Scruggs, T. E., Mastropieri, M. A., Forness, S. R., & Kavale, K. A. (1988). Early language intervention: A quantitative synthesis of single-subject research. *The Journal of Special Education*, *20*, 259-283.

*Seymour, F. W., & Stokes, T. F. (1976). Self-recording in training girls to increase work and evoke staff praise in an institution for offenders. *Journal of Applied Behavior Analysis*, *9*, 41-54.

*Sowers, J. S., Verrdi, M., Bourbeau, P., & Sheehan, M. (1985). Teaching job independence and flexibility to mentally retarded students through the use of a self-control package. *Journal of Applied Behavior Analysis*, *18*, 81-85.

Stage, S. A., & Quiroz, D. R. (1997). A meta-analysis of interventions to decrease disruptive classroom behavior in public education settings. *School Psychology Review*, *26*, 333-369.

Stahmer, A. C., & Schreibman, L. (1992). Teaching children with autism appropriate play in unsupervised environments using a self-management treatment package. *Journal of Applied Behavior Analysis*, *25*, 447-459.

*Stevenson, H. C., & Fantuzzo, J. W. (1984). Application of the "generalization map" to a self-control intervention with school-aged children. *Journal of Applied Behavior Analysis*, *17*, 203-212.

*Stevenson, H. C., & Fantuzzo, J. W. (1986). The generality and social validity of a competency-based self-control training intervention for underachieving students. *Journal of Applied Behavior Analysis*, *19*, 269-276.

Swanson, L. (1981). Modification of comprehension deficits in learning disabled children. *Learning Disability Quarterly*, *4*, 189-201.

Swanson, H. L., & Sachse-Lee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, *33*, 114-136.

*Trammel, D. L., Schloss, P. J., & Alper, S. (1994). Using self-recording, evaluation, and graphing to increase completion of homework assignments. *Journal of Learning Disabilities*, *27*, 75-81.

Wampold, B., & Worsham, N. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment*, *8*, 135-143.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual-subject research. *Behavioral Assessment*, *11*, 281-296.

Wilson, G. T., Leaf, R. C., & Nathan, P. E. (1975). The aversive control of excessive alcohol consumption by chronic alcoholics in the laboratory setting. *Journal of Applied Behavior Analysis*, *8*, 13-26.

*Wood, R., & Flynn, J. M. (1978). A self-evaluation token system versus an external evaluation token system alone in a residential setting with predelinquent youth. *Journal of Applied Behavior Analysis*, *11*, 503-512.

Wood, S. J., Murdock, J. Y., & Cronin, M. E. (2002). Self-monitoring and at-risk middle school students. *Behavior Modification*, *26*, 605-626.

Hsen-Hsing Ma was born on September 21, 1944. He graduated from the Department of Education, National Chengchi University, Taiwan and got his doctorate degree at the University Düsseldorf, Germany. The title of his dissertation was "The Experimental Single Subject Research in Education." Since 1980, he has taught behavior modification and research methodology in education. His important publications are as follows: Single Case Experimental Designs, *(1980);* Theories and Techniques in Behavior Modification *(5. ed.)(1990);* Research Methods in Educational Science *(1999; this book contains univariate and multivariate analysis and time-series analysis); "Meta-analysis of Criminal Theories in Terms of Social Evolutionism and Behavior Reinforcement Theory," (2002),* Journal of Education & Psychology, *25; and* Introduction to Educational Science *(2002; this book made a broad review of results of meta-analysis of between group empirical studies and tried to form the basis of knowledge body of repeatable educational science).*