Calculating Effect Size Estimates for Single-Subject Experimental Designs

The Eighth Annual International Campbell Collaboration Colloquium

Vancouver, British Columbia, Canada
May 12-14, 2008

Why Effect Sizes for Single-Subject Experiments?

- Single-subject experimental designs (SSEDs) traditionally evaluated by visual analysis
  - Evidence-based practice (EBP) emphasizes importance of more objective outcome measures, especially “magnitude of effect” indices or “effect sizes” (ES) (Horner et al., 2005).
- Meta-analysis: ES are needed to summarize outcomes from SSEDs for research synthesis
  - SSEDs predominant with low-incidence populations
  - ⇨ Meta-analysis of SSEDs emphasized in evidence hierarchies

Levels of Evidence

1. Meta-analysis of (a) single-subject experimental designs, (b) quasi-experimental group designs (i.e., non-randomized)		
2a. quasi-experimental *group designs**	2b. *Single-subject* experimental design – one intervention	2c. *Single-subject* experimental design – multiple interventions
3. Quantitative reviews that are non meta-analytic		
4. Narrative reviews		
5. Pre-experimental group designs and qualitative case studies		
6. Respectable opinion and/or anectodal evidence		

*** Consider differences in quality regarding threats to internal and external validity (Campbell & Stanley, 1963). Adapted from Schlosser, 2003.**

(Wendt & Lloyd, 2005)

Current Status

- What “effect size metrics” are most appropriate to measure effect size while respecting the characteristics of SSEDs
- Regression-based approaches
  - Piece-wise regression procedure (Center, Skiba, & Casey, 1986)
  - 4-parameter model (Huitema & McKean, 2000)
  - Multilevel Models (Van den Noortgarte & Onghena, 2003a, 2003b)
- Non-regression-based approaches
  - Family of “non-overlap” metrics
  - Specific metrics for behavior increase versus behavior reduction data

Program

- *A Review on the Application of the Percentage of Non-overlapping Data (PND) in Behavioral Science and Special Education Research*

  Presenter: Ralf W. Schlosser, Ph.D., Northeastern University, U.S.A.

- *Computation of Non-regression-based Effect Size Metrics*

  Presenter: Oliver Wendt, Ph.D., Purdue University, U.S.A.

- *Computation of Regression-based Effect Size Metrics*

  Presenter: Tasha Beretvas, Ph.D., University of Texas at Austin, U.S.A.

Computation of Non-regression-based Effect Size Metrics

Oliver Wendt, Ph.D.

Purdue University

U.S.A.

General Considerations and Precautions

- Requirements of the specific metric applied
  - Minimum number of data points or participants
  - Specific type of SSED (e.g., multiple baseline)
  - Randomization (e.g., assignment to treatment conditions, order of participants)
  - Assumptions of data distribution and nature (e.g., normal distribution, no autocorrelation)
- Limitations
  - Ability to detect changes in level and trend
  - Orthogonal slope changes
  - Sensitivity to floor and ceiling effects
- Direction of behavior change
  - Behavior increase vs. decrease

Non-regression Based Approaches

- Family of “non-overlap” metrics
  - Percentage of Non-overlapping Data (PND)
  - Percentage of All Non-overlapping Data (PAND)
  - Percentage of Data Points Exceeding the Median (PEM)
  - Pairwise Data Overlap (PDO)

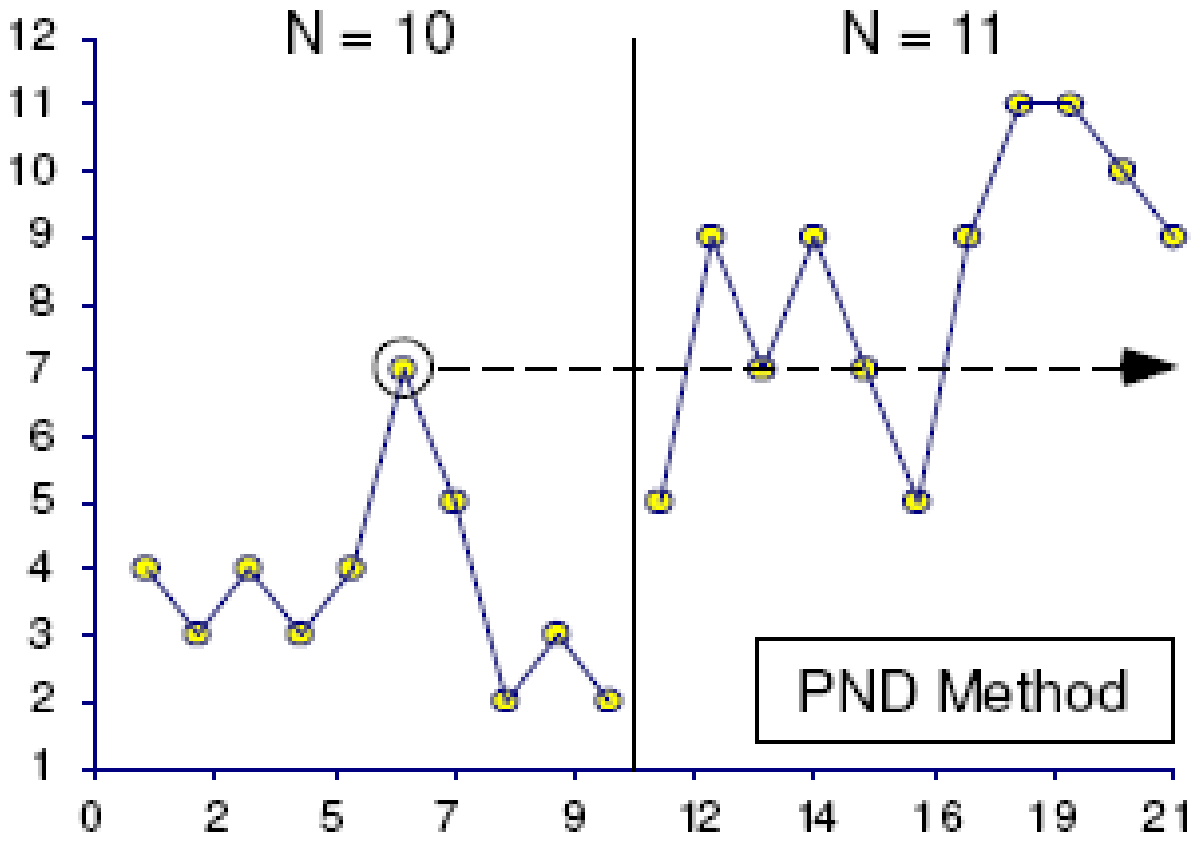
- Specific metrics for behavior reduction data (vs. behavior increase data)
  - Percentage Reduction Data (PRD)
  - Percentage of Zero Data (PZD)

Percentage of Nonoverlapping Data (PND)

- Calculation of non-overlap between baseline and successive intervention phases (Scruggs, Mastropieri, & Casto, 1987)
  - Identify highest data point in baseline and determine the percentage of data points during intervention exceeding this level
- easy to interpret
- non-parametric statistic

PND Calculation: An Example

PND = 7/11 = 60%



Interpretation of PND Scores

- If a study includes several experiments, PND scores are aggregated by taking the median (rather than mean)
  - Scores usually not distributed normally
  - Median less effected by “outliers”
- PND statistic: the higher the percentage the more effective the treatment
- Specific criteria for interpreting PND scores outlined by Scruggs, Mastropieri, Cook, and Escobar (1986)

Interpretation of PND Scores (cont.)

- PND range 0-100%
  - PND < 50% reflects unreliable treatment
  - PND 50% - 70% questionable effectiveness
  - PND 70% - 90% fairly effective
  - PND > 90% highly effective

Limitations of PND

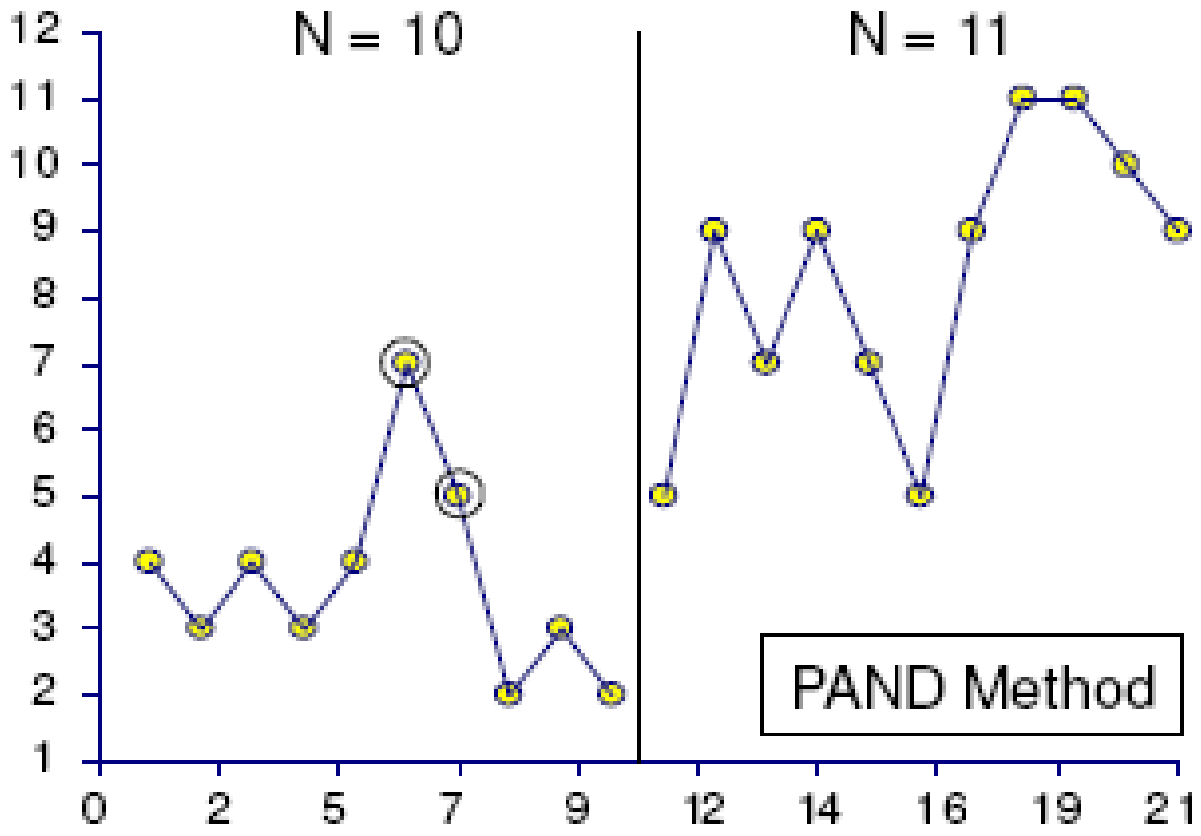
- Ignores all baseline data except one datapoint (this one can be unreliable)
  - Ceiling effects
- Lacks sensitivity or discrimination ability as it nears 100% for very successful interventions
  - Can not detect slope changes
- Needs its own interpretation guidelines
  - Technically not an effect size

Percentage of All Non-Overlapping Data (PAND)

- Calculation of total number of data points that do not overlap between baseline and intervention phases (Parker, Hagan-Burke, & Vannest, 2007)
  - Identify overlapping data points (minimum number that would have to be transferred across phases for complete data separation)
  - Compute % overlap by dividing number of overlapping points by total number of points
  - Subtract this percent from 100 to get PAND
- Non-parametric statistic

PAND Calculation: An Example

% Overlap = 2/21 = 9.5%
PAND = 100%-9.5%= 90.5%



Advantages of PAND

- Uses all data points across both phases
- May be translated to *Phi* and *Phi*^{2} to determine effect size (e.g., Cohen's *d*)

Limitations of PAND

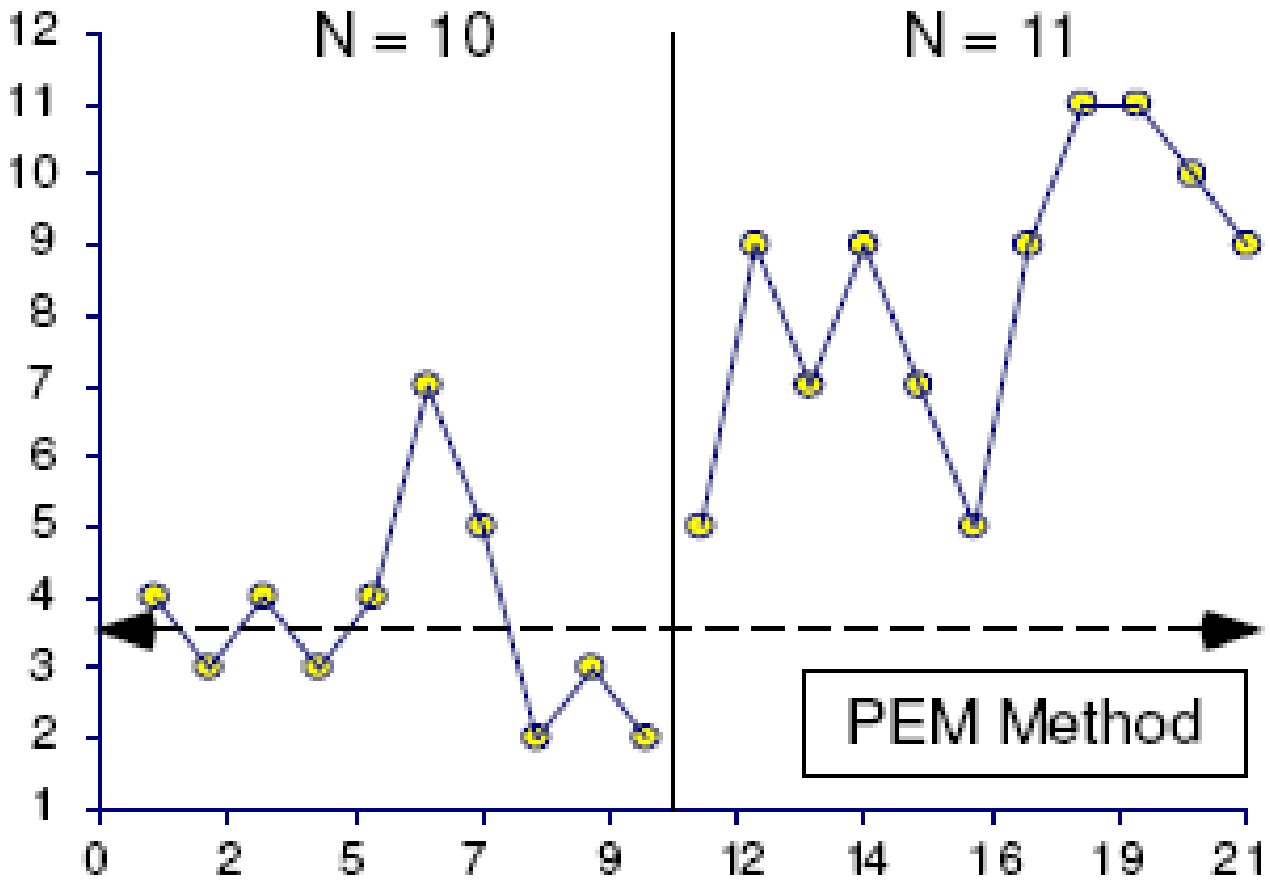
- Insensitive at the upper end of the scale
  - 100% is awarded regardless of distance between data points in the two phases
- Measures only mean level shifts and does not control for positive baseline trend
- Requires 20 data points for calculation

Percentage of Data Points Exceeding the Median (PEM)

- Calculation of percentage of data points exceeding the median of baseline phase (Ma, 2006)
  - Locate median point (uneven data set) or point between the two median points (even data set) in baseline data
  - Draw horizontal middle line passing through median of baseline into treatment phase
  - Compute percentage of treatment phase data points above the middle line if behavior increase is expected
    - Below middle line if behavior decrease is expected
- Non-parametric statistic

PEM Calculation: An Example

PEM =11/11 = 100%



Interpretation of PEM

- Null hypothesis
  - If treatment is ineffective, data points will continually fluctuate around the middle line
- PEM scores range from 0 to 1
  - .9 to 1 reflects highly effective treatment
  - .7 to .9 reflects moderately effective treatment
  - Less than .7 reflects questionable or not effective treatment

Advantages of PEM

- PEM scores can be calculated from each pair of baseline-treatment phases
- Mean effect size of each article or each variable category can be calculated
- In the presence of floor or ceiling data points, PEM still reflects effect size while PND does not

Limitations of PEM

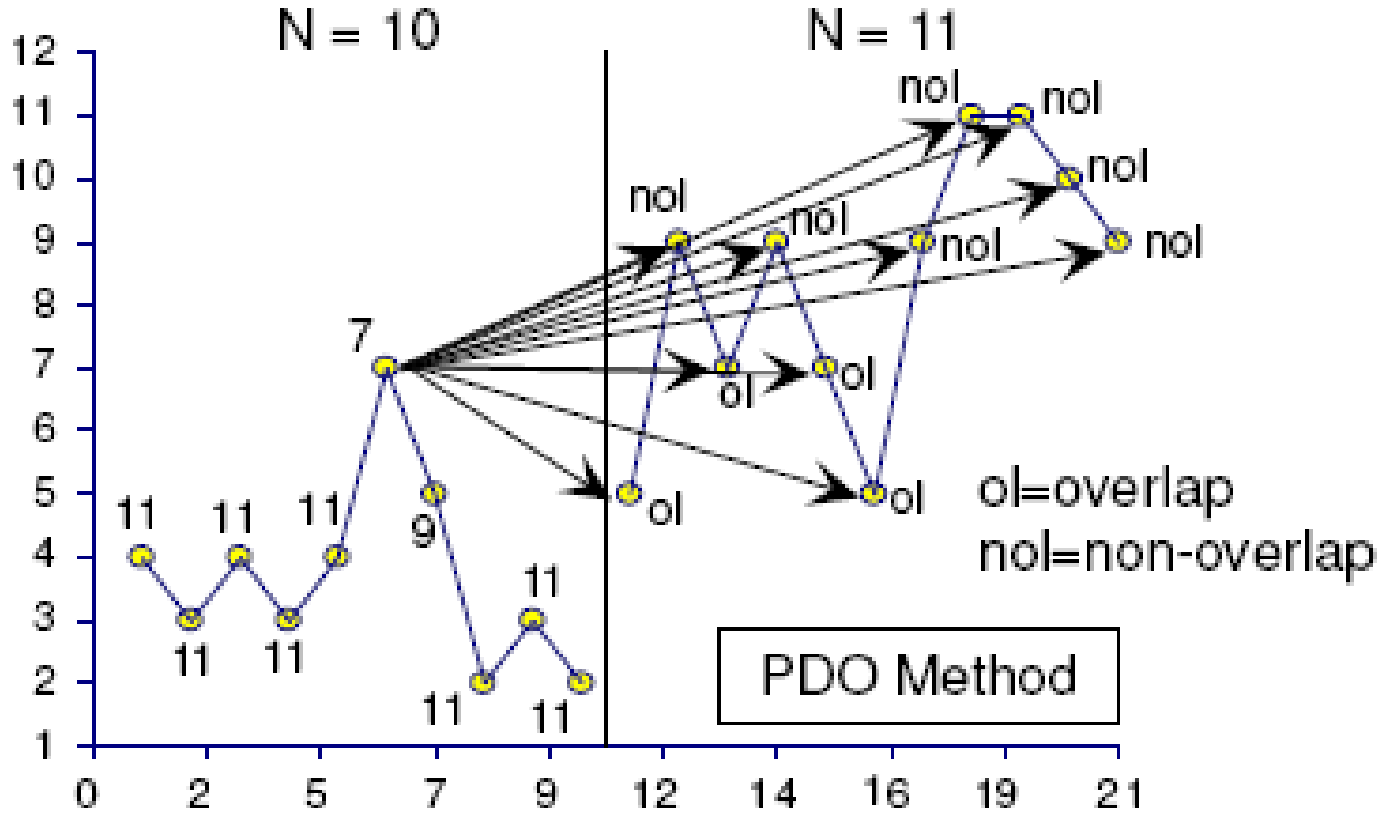
- Insensitive to magnitude of data points above the median
- Does not consider trend and variability in data points of treatment phase
- May reflect only partial improvement if orthogonal slope is present in baseline treatment pair after first treatment phase

Pairwise Data Overlap (PDO)

- Calculation of overlap of all possible paired data comparisons between baseline and intervention phases (Parker & Vannest, in press)
  - Compare baseline data point with all intervention data points
  - Determine number of overlapping (ol) and non-overlapping (nol) points
  - Compute total number of “nol” points divided by total number of comparisons
- Non-parametric statistic

PDO Calculation: An Example

PDO = (11 +11 + 11 + 11 + 11 + 7 + 9 + 11 + 11 + 11) / (10 x 11) = 95%



Advantages of PDO

- Produces highly reliable results as indicated by narrow confidence intervals
- Relates closely to established effect sizes (Pearson *R*, Kruskal-Wallis *W*)
- Matches visual judgments as well as other indices

Limitations of PDO

- Takes slightly longer to calculate
- Requires that individual data point results be written down and added
- Calculation is laborious for long and crowded data series

How Do PND, PEM, and PAND Compare?

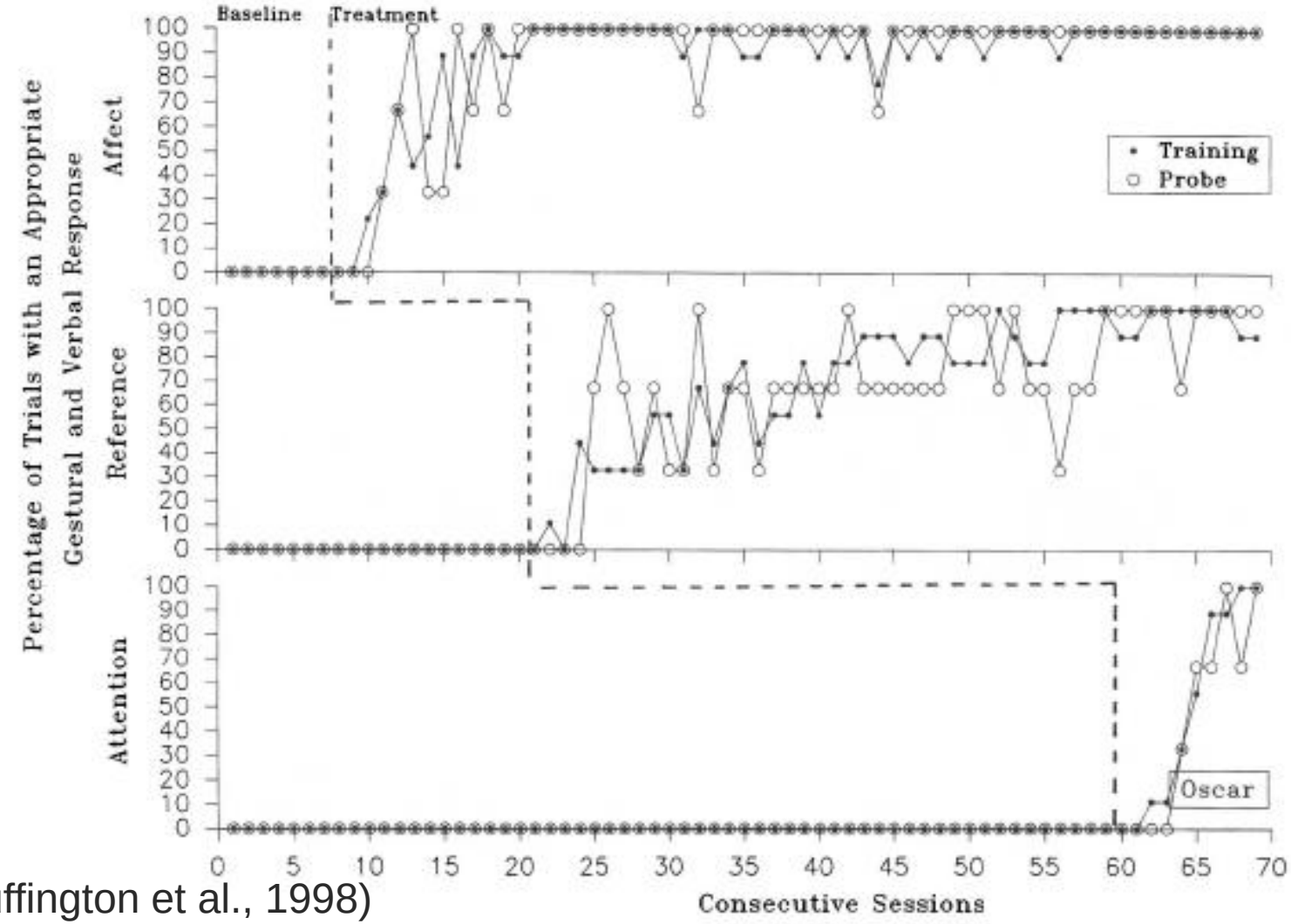
- All three effect size metrics were applied to “real data”
- Data set taken from systematic review of school-based instructional interventions for students with autism spectrum disorders (Machalicek et al., in press)
- Outcomes: communication skills (e.g., gestures, natural speech, use of speech-generating device)
- N=9 studies, 24 participants, various designs, A-B phases (acquisition) extracted

Study/ Design	Purpose	Participants	Outcomes	PND (%)	PEM (%)	PAND (%)
Johnson et al. (2004)/ MBD	Evaluate embedded instruction	A1 A2 A3	% correct use of AAC device	100.00 85.71 87.50	100.00 100.00 100.00	100.00 87.50 95.83
Buffington et al. (1998)/ MPD across behaviors	To demonstrate the effects of modeling, prompting, & reinforcement on gesture acquisition	A1 (1st response) A2 (affect) A3 (reference) B1 B2 B3 C1 C2 C3 D1 D2 D3	% trials appropriate gestural and verbal response *NOTE ZERO BASELINES ⇨ PND and PEM were the same* *PAND better as it is not influenced by number of datapoints (with zero baselines) show example*	94.44 97.37 93.33 96.77 95.92 80.00 90.57 97.37 90.00 98.08 94.87 92.00	94.44 97.37 93.33 96.77 95.92 80.00 90.57 97.37 90.00 98.08 94.87 92.00	90.38 98.21 96.43 97.10 97.10 97.10 91.38 98.18 98.28 98.41 96.83 96.83
Hetzroni & Tannous (2004)/ MBD	To evaluate the effectiveness of using a computer-based intervention for enhancing communication functions	B4 (play) B5 (food) B6 (hygiene) C3 (play) C2 (food) C6 (hygiene) E1 (play) E2 (food) E3 (hygiene)	frequency of relevant speech frequency communication initiations	94.44 41.67 50.00 94.44 75.00 50.00 61.11 8.33 0.00	94.44 66.67 100.00 100.00 83.33 66.67 61.11 16.66 33.33	95.45 68.18 90.91 95.45 86.33 86.33 68.18 54.55 00.00

Study/ Design	Purpose	Participants	Outcomes	PND (%)	PEM (%)	PAND (%)
Schepis et al. (1998)/ MPD	Evaluate effects of naturalistic teaching strategies and SGD use	A1 (snack) A2 (play) B1 (snack) B2 (play) C D	frequency communicative interactions per min. *all 3 metrics no problem when clear phase separation,*	100.00 100.00 100.00 100.00 100.00 100.00	100.00 100.00 100.00 100.00 100.00 100.00	100.00 100.00 100.00 100.00 100.00 100.00
Hetzroni and Shalem (2005)/ MPD across subjects	Evaluate effectiveness of multilevel software in teaching orthography	A1 B1 C1 A2 B2 C2	# of correctly matched orthographic symbols to corresponding logos *Outlier distort PND*	66.67 70.00 87.50 60.00 81.82 50.00	77.77 100.00 100.00 70.00 100.00 100.00	78.57 83.33 95.00 73.33 90.00 80.95
Sigafoos et al. (2004)/ MBD	Repair comm. breakdowns	A	% correct for blocks of 3 opportunities to locate AAC device	100.00	100.00	100.00

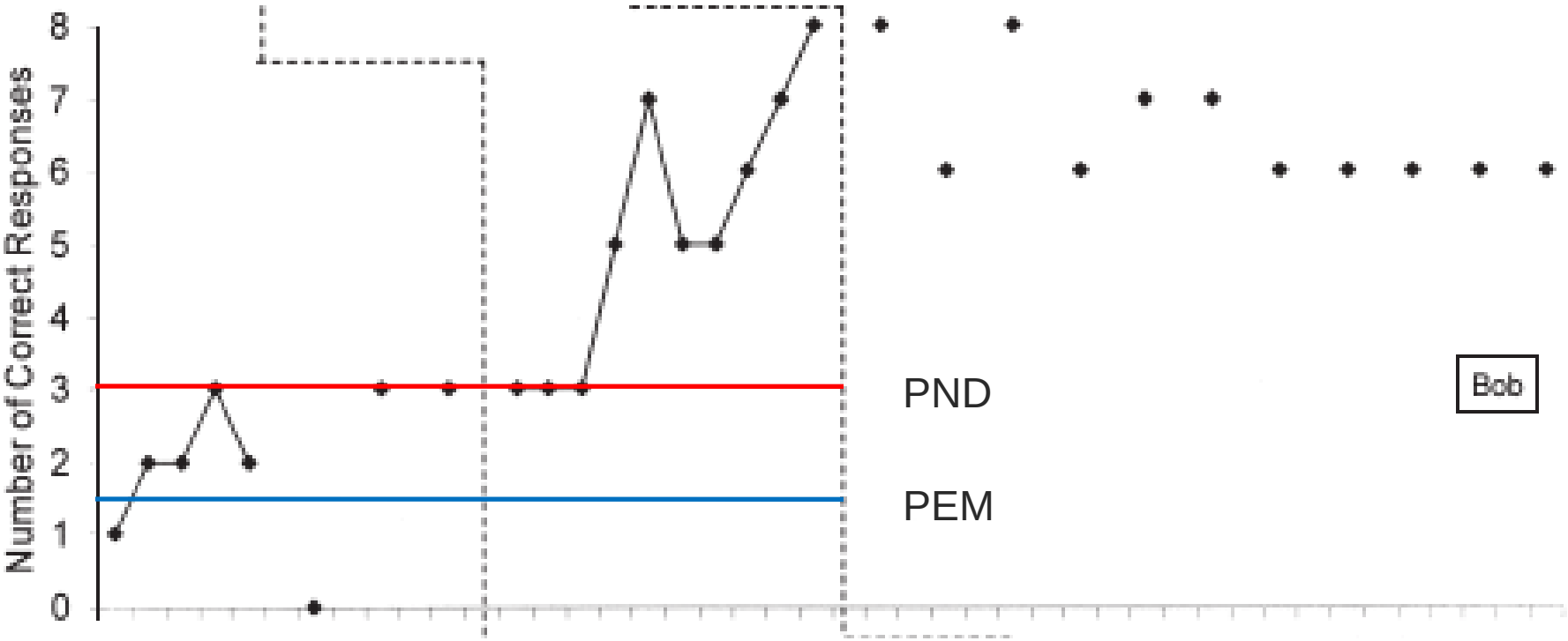
Study/ Design	Purpose	Participants	Outcomes	PND (%)	PEM (%)	PAND (%)
Keen et al. (2001)/ MPD across behaviors	To evaluate the effectiveness of a teacher-implemented intervention intended to substitute prelinguistic behaviors with functional communication	A1 (greet –prel.)	% of opportunities	00.00	60.00	00.00
		A2 (request - pre.)		83.33	100.00	91.67
		A3 (choice-pre.)		100.00	100.00	100.00
		A4 (greet – repl.)		60.00	70.00	85.71
		A5 (request – rep.		50.00	70.00	91.67
		A6 (choice-repl.)		100.00	100.00	100.00
		B1		81.25	87.50	81.25
		B2		100.00	100.00	100.00
		B3		80.00	100.00	90.48
		B4		88.24	88.24	87.50
		B5		81.81	81.81	91.67
		B6		83.33	83.33	90.48
		C1		0.00	38.88	39.13
		C2		63.64	90.91	61.10
		C3		66.67	66.67	92.31
		C4		00.00	83.33	78.26
		C5		54.55	72.73	72.22
		C6		83.33	83.33	92.31
		D1		00.00	00.00	00.00
		D2		92.86	100.00	94.12
		D3		0.00	0.00	71.43
		D4		0.00	0.00	00.00
		D5		85.71	85.71	88.24
		D6		100.00	100.00	100.00
Smith & Camarata (1999)/MBD	evaluate natur. language teaching interv.	A	% intervals verbal interaction	40.00	40.00	83.33
		B		28.57	100.00	92.59
		C		43.48	100.00	87.50
Kravits et al. (2002)/MBD	examine PECS effects	A1 (centers)		91.67	91.67	96.55
		A2 (journals)		90.00	90.00	97.14

Observed Patterns: PEM/PND versus PAND



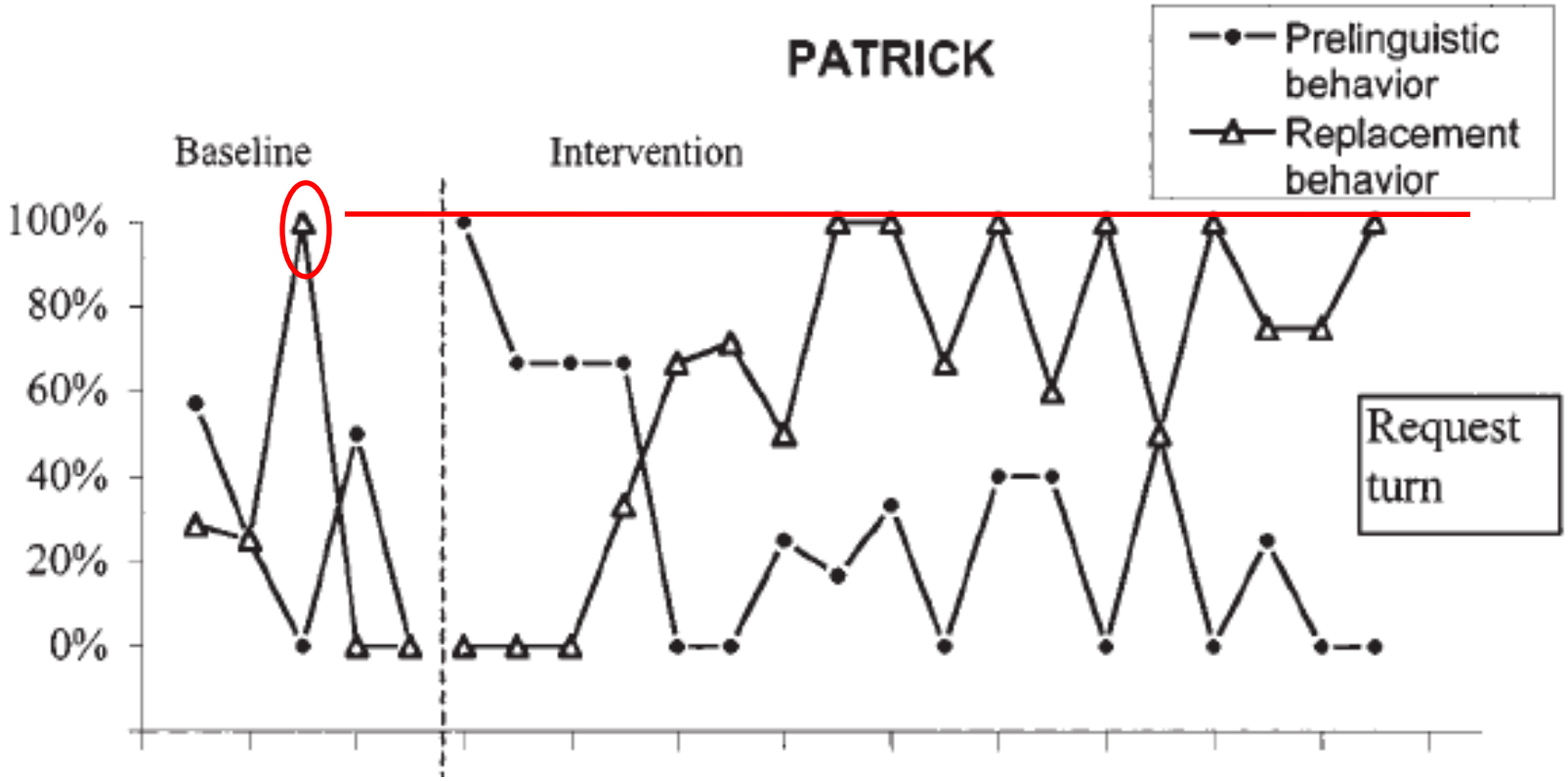
(Buffington et al., 1998)

Observed Patterns: PEM Inflated Scores



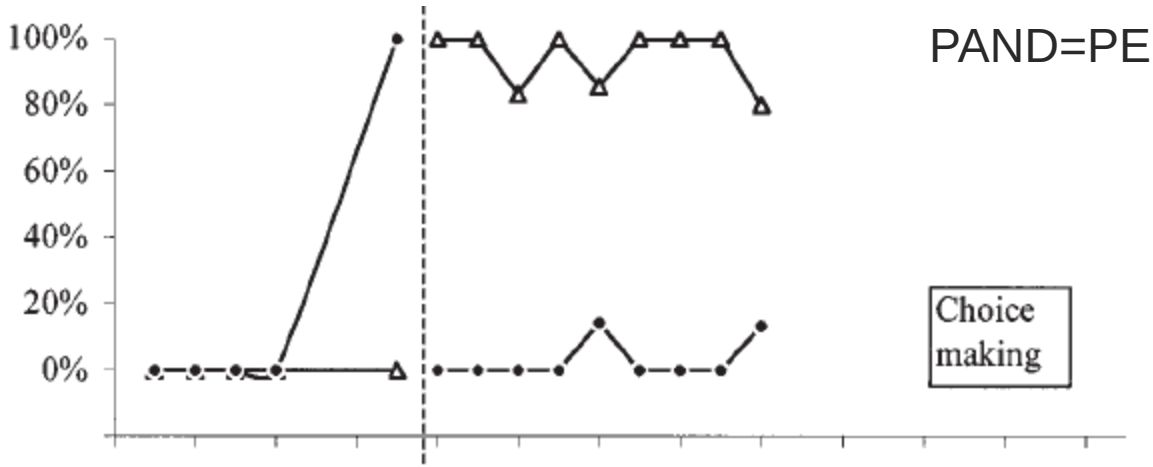
Hetzroni & Shalem, 2005

Observed Patterns: PND Distorted by Outliers

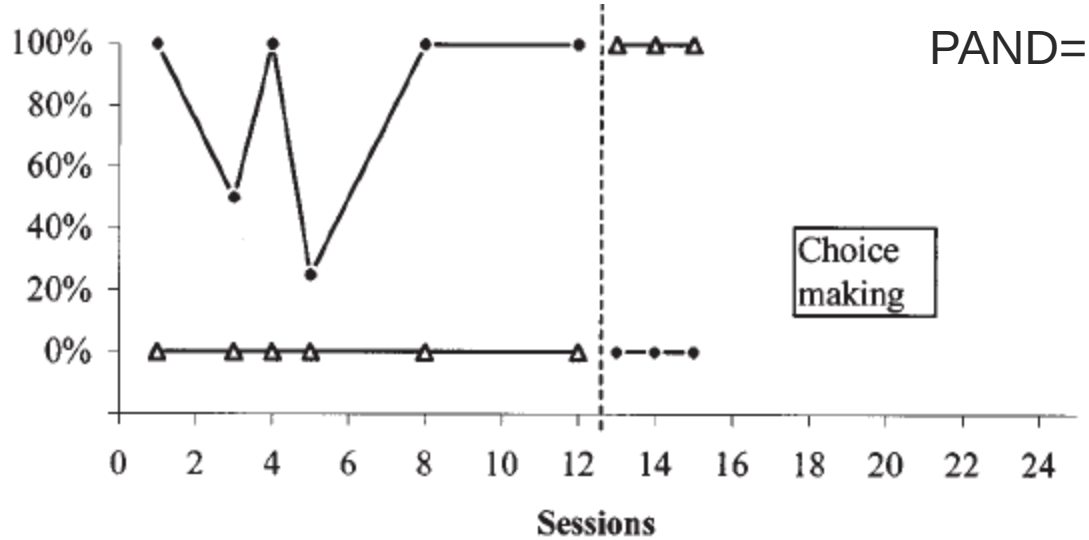


Keen, Sigafoos, & Woodyatt, 2001

Observed Patterns: PAND, PEM, PND Insensitivity

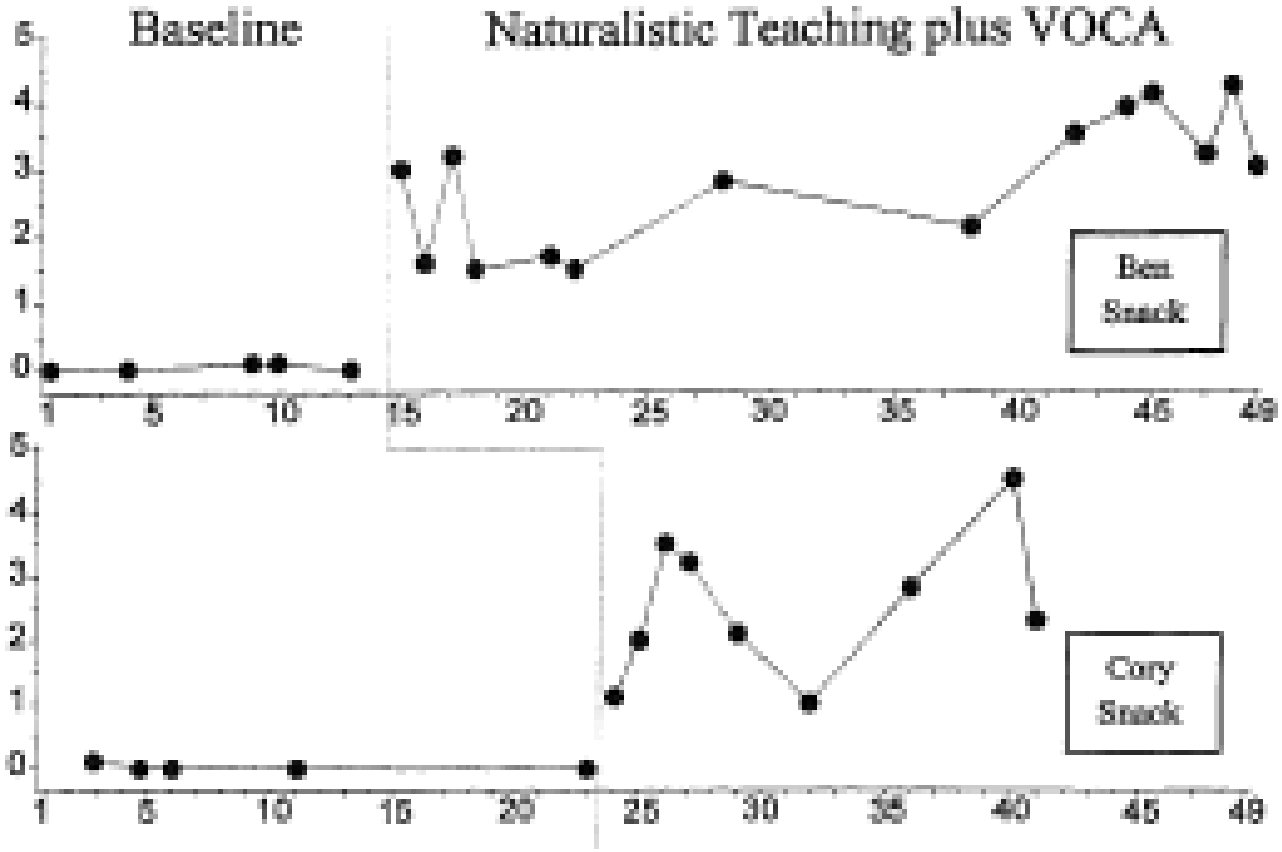


PAND=PEM=PND= 100%



PAND=PEM=PND= 100%

Observed Patterns: PAND, PEM and PND w/out Overlap

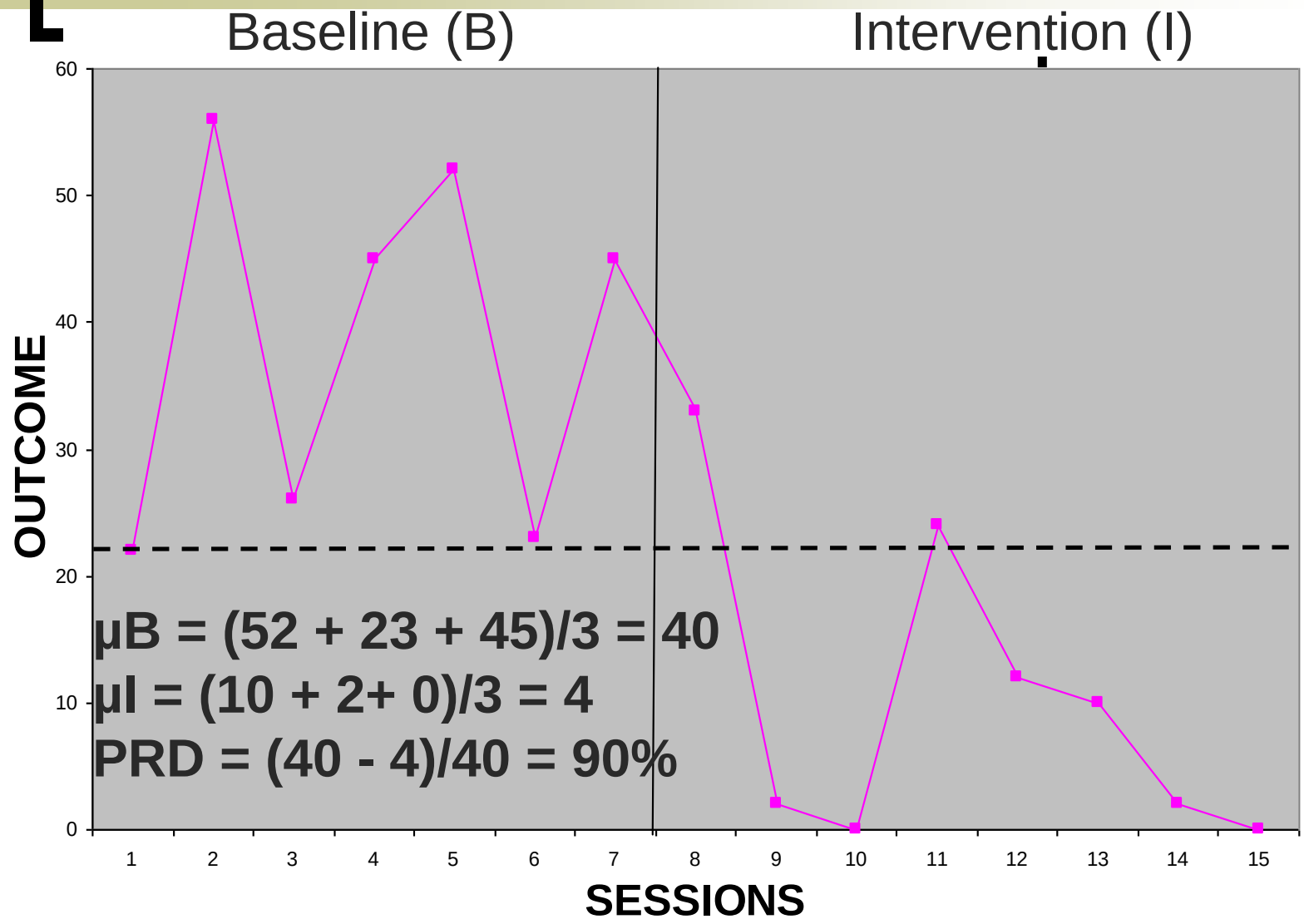


(Schepis et al., 1998)

Percentage Reduction Data (PRD)

- Calculation of reduction of targeted behavior due to intervention (O'Brien & Repp, 1990)
  - Determine mean of last three data points from baseline (μB) and of last three data points from intervention (μI)
  - Calculate the amount of change between baseline and treatment
    - $[(\mu B - \mu I) \div \mu B] \times 100$
- Also called Mean Baseline Reduction (MBR) (Campbell, 2003, 2004)

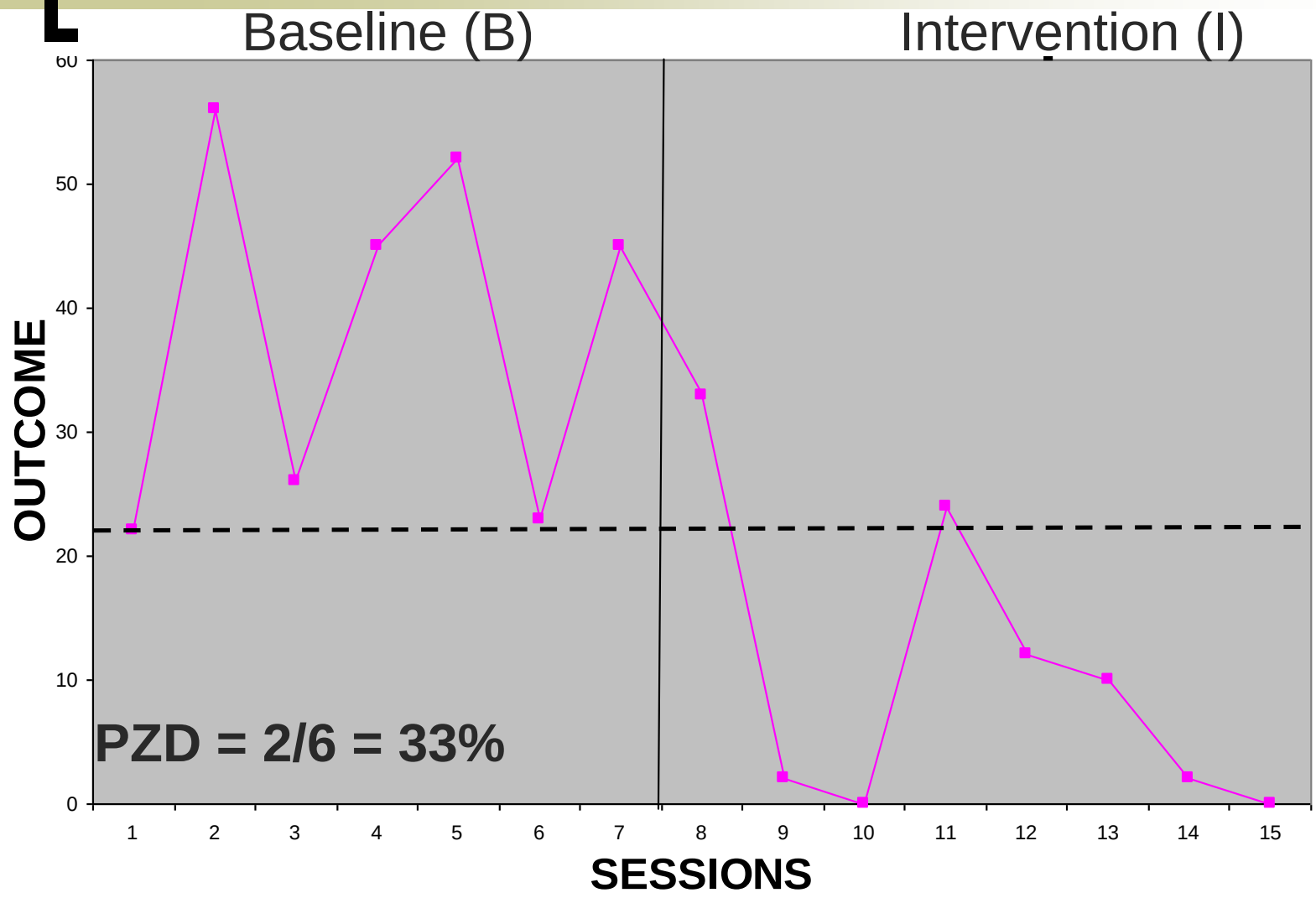
PRD Calculation: An Example



Percentage of Zero Data (PZD)

- Calculation of the degree to which intervention completely suppresses targeted behavior (Scotti et al., 1991)
  - Identify first data point to reach zero in an intervention phase
  - Calculate the percentage of data points that remain at zero from the first zero point onwards

PZD Calculation: An Example



Interpretation of PZD scores

- PZD range 0-100%
  - PZD < 18% reflects ineffectiveness
  - PZD 18% - 54% reflects questionable effectiveness
  - PZD 55% - 80% reflects fair effectiveness
  - PZD > 80% reflects high effectiveness

How Do PND, PRD, and PZD Compare?

- All three effect size metrics were applied to “real data”
- Data set taken from systematic review of intervention studies on Functional Communication training (Chacon, Wendt, & Lloyd, 2006)
- Intervention: graphic symbols, manual signs, speech

Study	N	Condition	Intervention	PND (%)	PRD (%)	PZD (%)
Braithwaite & Richdale (2000)	1	Access	Vocalizations (“I want ___ please”)	100	100	95
		Escape	Vocalizations (“I need help please”)	92	100	100
Day, Horner, & O’Neill (1994)	1	Access	Manual signs (“want”)	100	100	67
		Escape	Vocalizations (“go”)	100	98	50
Horner & Day (1991)	1	Escape	Graphic symbol (“break”) with 1s delay	100	83	80
			Graphic symbol (“break”) with 20s delay	0	-64	0
Schindler & Horner (2005)	2	Escape	Graphic symbol/Vocalization (“help”)	18	83	47
		Escape	Graphic symbol (“change,” “activity”)	72	92	24
Sigafoos & Meikle (1996)	2	Attention	Vocalizations (“Beth”)	86	81	80
		Access	Vocalizations (“drink,” “toy,” “want”)	100	100	96
		Attention	Gestures (tapping teacher’s hand)	100	100	100
		Access	Graphic symbols (“food,” “drink,” “toy”)	100	100	100
Wacker et al. (1990)	1	Access	Gestures (touching chin)	50	49	25

Future Directions

- Currently no consensus supporting the use of a single “best” effect size metric for summarizing the results of SSEDS
- Future meta-analyses may use several SSED effect sizes to summarize a treatment’s effectiveness
  - Use of these multiple measures should provide a better picture describing how different metrics compare to one another and to fit each single-subject research design with the most suitable metric

Future Directions (cont.)

- Future research on SSED effect size metrics needs to compare the metrics with each other using real datasets and not a convenience sample
  - Compare metrics within family of “non-overlap”
  - Compare across regression versus non-regression based approaches

Contact Information

Oliver Wendt, Ph.D.
Purdue AAC Program
Department of Educational Studies
BRNG 5108, Purdue University
West Lafayette, IN 47907-2098, USA
Phone: (+1) 765-496-8314
Fax: (+1) 765-496-1228
E-mail: olli@purdue.edu
http://www.edst.purdue.edu/aac

References

- Braithwaite, K. L., & Richdale, A. L. (2000). Functional communication training to replace challenging behaviors across two behavioral outcomes. *Behavioral Interventions, 15*, 21-36.
- Buffington, D. M., Krantz, P. J., McClannahan, L. E., & Poulson, C. L. (1998). Procedures for teaching appropriate gestural communication skills to children with autism. *Journal of Autism and Developmental Disorders, 28*(6), 535-545.
- Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24,* 120-138.
- Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*(2), 234-246.
- Campbell, J. M. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: RandMcNally.
- Center, B. A., Skiba, R. J., & Casey, A. (1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*(4), 387-400.

References (cont.)

- Chacon, M., & Wendt, O., & Lloyd, L.L. (2006, November). *Using Functional Communication Training to reduce aggressive behavior in autism.* Paper presented at the Annual Convention of the American Speech-Language-Hearing Association (ASHA), Miami, FL.
- Cooper, H. M., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.
- Day, H. M., Horner, R. H., & O'Neill, R. E. (1994). Multiple functions of problem behaviors: Assessment and intervention. *Journal of Applied Behavior Analysis, 27*, 279-289.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.
- Hetzroni, O. E., & Shalem, U. (2005). From logos to orthographic symbols: A multilevel fading computer program for teaching nonverbal children with autism. *Focus on Autism and Other Developmental Disabilities, 20*(4), 201-212.
- Hetzroni, O., & Tannous, J. (2004). Effects of a computer-based intervention program on the communicative functions of children with autism. *Journal of Autism and Developmental Disorders, 34*(2), 95-113.

References (cont.)

- Horner et al. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165-179.
- Horner, R. H., & Day, H. M. (1991). The effects of response efficiency on functionally equivalent competing behaviors. *Journal of Applied Behavior Analysis, 24*, 719-

  732.
- Huitma, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38-58.
- Johnson, J. W., McDonnell, J., Holzwarth, V. N., & Hunter, K. (2004). The efficacy of embedded instruction for students with developmental disabilities enrolled in general education classes. *Journal of Positive Behavior Interventions, 6*(4), 214-227.
- Keen, D., Sigafoos, J., & Woodyatt, G. (2001). Replacing prelinguistic behaviors with functional communication. *Journal of Autism and Developmental Disorders, 31*(4), 385-398.

References (cont.)

- Kravits, T. R., Kamps, D. M., Kemmerer, K., & Potucek, J. (2002). Brief report: Increasing communication skills for an elementary-aged student with autism using the picture exchange communication system. *Journal of Autism and Developmental Disorders, 32,* 225-230.
- Ma, H. (2006). An alternative method for quantitative synthesis of single-subject researchers: Percentage of data points exceeding the median. *Behavior Modification, 30*(5), 598-617.
- Machalicek et al. (in press.). A review of school-based instructional interventions for students with autism spectrum disorders. *Research in Autism Spectrum Disorders.*
- O'Brien, S., & Repp, A. C. (1990). Reinforcement-based reductive procedures: A review of 20 years of their use with persons with severe or profound retardation. *Journal of the Association for Persons with Severe Handicaps, 15,* 148-159.
- Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*(4), 194-204.

References (cont.)

- Parker, R. I. & Vannest, K. J. (in press). Pairwise data overlap for single case research. *School Psychology Review.*
- Pennington, L. (2005). Book Review. *International Journal of Language and Communication Disorders, 40*(1), 99-102.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.
- Schepis, M. M., Reid, D. H., Behrmann, M. M., & Sutton, K. A. (1998). Increasing communicative interactions of young children with autism using a voice output communication aid and naturalistic teaching. *Journal of Applied Behavior Analysis, 31,* 561-578.
- Schindler, H. R., & Horner, R. H. (2005). Generalized reduction of problem behavior of young children with autism: Building trans-situational interventions. *American Journal on Mental Retardation, 110*, 36-47.
- Schlosser, R. W. (2003). *The efficacy of augmentative and alternative communication: Toward evidence-based practice.* San Diego, CA: Academic Press.

References (cont.)

- Schlosser, R. W. (2005). Reply to Pennington: Meta-analysis of single-subject research: How should it be done? *International Journal of Language and Communication Disorders, 40*(3), 375-378.
- Scotti, J. R., Evans, I. M., & Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation, 96*(3), 233-256.
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research methodology: Methodology and validation. *Remedial and Special Education, 8,* 24-33.
- Scruggs, T. E., Mastropieri, M. A., Cook, S. B., & Escobar, C. (1986). Early intervention for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders, 11,* 260-271.
- Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation method: Building the evidence base for method choice in government sponsored evaluation* (pp. 95-109). San Francisco: Jossey-Bass.

References (cont.)

- Sigafoos, J., & Meikle, B. (1996). Functional communication training for the treatment of multiply determined challenging behavior in two boys with autism. *Behavior Modification, 20*(1), 60-84.
- Sigafoos, J., O'Reilly, M., Seely-York, S., & Chaturi, E. (2004). Teaching students with developmental disabilities to locate their AAC device. *Research in Developmental Disabilities, 25,* 371-383.
- Smith, A. E., & Camarata, S. (1999). Using teacher-implemented instruction to increase language intelligibility of children with autism. *Journal of Positive Behavior Interventions, 1*(3), 141-151.
- Van den Noortgarte, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325-346.

References (cont.)

- Van den Noortgarte, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, and Computers, 35*(1), 1-10.
- Wacker, D. P., Steege, M. W., Northup, J., Sasso, G., Berg, W., Reimers, T., Cooper, L., Cigrand, K., & Donn, L. (1990). A component analysis of functional communication training across three topographies of severe behavior problems. *Journal of Applied Behavior Analysis, 23,* 417-429.
- Wendt, O., & Lloyd, L.L. (2005). *Evidence-based practice in AAC.* Unpublished PowerPoint Presentation, Purdue University, West Lafayette.