

Use of a Direct Observational Measure in a Trial of Risperidone and Parent Training in Children with Pervasive Developmental Disorders

Benjamin L. Handen · Cynthia R. Johnson · Eric M. Butter · Luc Lecavalier · Lawrence Scahill · Michael G. Aman · Christopher J. McDougle · L. Eugene Arnold · Naomi B. Swiezy · Denis G. Sukhodolsky · James A. Mulick · Susan W. White · Karen Bearss · Jill A. Hollway · Kimberly A. Stigler · James Dziura · Sunkyung Yu · Kelley Sacco · Benedetto Vitiello

Published online: 20 November 2012

© Springer Science+Business Media New York 2012

Abstract A Structured Observational Analog Procedure (SOAP), an analogue measure of parent-child interactions, was used to assess treatment outcome in children with Autism Spectrum Disorder and serious behavior problems. It served as a secondary

Acknowledgement of funding Supported by the following cooperative agreement grants from the National Institute of Mental Health (NIMH): U10MH66768 (P.I.: M. Aman), U10MH66766 (P.I.: C. McDougle), and U10MH66764 (P.I.: L. Scahill).

Trial Registration: NCT00080145

B. L. Handen · C. R. Johnson
University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

E. M. Butter · L. Lecavalier · M. G. Aman · L. E. Arnold · J. A. Mulick · J. A. Hollway
Ohio State University, Columbus, OH, USA

L. Scahill · K. Bearss · J. Dziura
Emory University School of Medicine, Atlanta, GA, USA

C. J. McDougle
Harvard Medical School, Boston, MA, USA

N. B. Swiezy · K. A. Stigler
Indiana University School of Medicine, Indiana, IN, USA

D. G. Sukhodolsky
Yale University, New Haven, CT, USA

S. W. White
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

outcome measure in a 24-week, randomized trial of risperidone (MED; $N=49$) versus risperidone plus parent training (COMB; $n=75$) (ages 4–13 years). At 24-weeks, there was 28 % reduction in child inappropriate behavior during a Demand Condition ($p=.0002$) and 12 % increase in compliance to parental requests ($p=.004$) for the two treatment conditions combined. Parents displayed 64 % greater use of positive reinforcement ($p=.001$) and fewer repeated requests for compliance ($p<.0001$). In the analysis of covariance (ANCOVA), COMB parents used significantly more positive reinforcement ($p=.01$) and fewer restrictive statements ($p<.05$) than MED parents. The SOAP is sensitive to change in child and parent behavior as a function of risperidone alone and in combination with PMT and can serve as a valuable complement to parent and clinician-based measures.

Keywords Autism spectrum disorder · Autism · Observational measures · Behavioral interventions · Parent training · Clinical trials

Psychopharmacological trials in children with developmental disabilities, including Autism Spectrum Disorder (ASD), have often used parent and/or teacher behavior problem checklists to assess efficacy. These rating tools are dimensional measures that often include a set of behaviors that are scored on a Likert scale (e.g., “no problem,” “mild problem,” “moderate problem,” or “severe problem”). Behavior problem checklists can assess a range of target behaviors, including both externalizing disorders (e.g., ADHD symptoms, ODD symptoms) and internalizing disorders (e.g., symptoms of anxiety or depression). Such measures have become the standard for documenting change in behavior in psychopharmacological trials in childhood behavioral disorders (MTA Cooperative Group 1999; RUPP Autism Network 2002; 2005). Behavior problem checklists offer a number of advantages to researchers. First, most are well-normed and are able to assess a wide range of domains. Second, they are fairly easy to administer and can be mailed to teachers for completion. Third, they have been found to be sensitive to drug effects. Fourth, they allow ratings to be averaged across time, including both good and bad days, so that the measure is not biased by one exceptional day that happens to occur at the time of measurement. Despite their ease of use and record of reliability, these tools also have some inherent weaknesses, including problems with the rater misunderstanding some items, rater bias, reluctance of some raters to “say anything bad” about a child, and especially

S. Yu
University of Michigan Medical School, Ann Arbor, MI, USA

K. Sacco
Children’s Hospital of Pittsburgh, Pittsburgh, PA, USA

B. Vitiello
National Institute of Mental Health, Bethesda, MD, USA

B. L. Handen (✉)
Merck Program, Western Psychiatry Institute and Clinic, 1011 Bingham St., Pittsburgh, PA 15203,
USA
e-mail: handenbl@upmc.edu

cross-rater reliability problems, given that each child usually has a separate rater, introducing inter-subject variability. Although the scales include instructions, there is usually no reliability training or calibration for the caregivers who fill them out.

Measures based on direct observation have long been a mainstay of intervention research in the fields of applied behavior analysis and child psychology (Cunningham and Barkley 1979; Webster-Stratton 1996; McMahon and Forehand 2003; Taylor and Hoch 2008). Using various formats, direct observational measures may involve assessment of behavior in real time or through examination of video recordings. For example, trained raters may score standardized scales based on observation in the classroom (Swanson et al. 1998). Alternatively, parents or teachers may compile “counts” of specific behaviors in school or in the home (Cooper et al. 2007). Within the ASD and developmental disabilities fields, direct observations have been used to assess treatment efficacy in a wide range of problem areas, from self-injury (Yang 2003), to increasing food acceptance (Ahearn 2002), to eliminating aggression (Kahng et al. 2008). Such measures have also been used to assess treatment efficacy in numerous psychopharmacology trials among a range of childhood disorders, especially ADHD (e.g., Abikoff et al. 2004; Pelham et al. 2002).

Within the psychopharmacology literature among children with ASD, there has been less of an emphasis on direct observation, as most studies have focused on global ratings and behavior problem checklists as the primary assessment measures (e.g., RUPP Autism Network 2002, 2005; King et al. 2009). Direct observation of mother-child interactions has been even rarer still. This is an important area of interest, as improvement in child behavior often influences change in parent behaviors. In addition, interventions that seek directly to impact how parents interact with their children also should use such observations to demonstrate change in parental behavior. In fact, direct observation has been used to document change in parent and child behavior in a small number of psychopharmacology trials (although not necessarily among children with ASD). For example, Barkley and Cunningham (1979) found hyperactive boys to be significantly more compliant during mother-child interactions when on methylphenidate compared to placebo or no drug. Handen et al. (1999) used an analogue observation to examine changes in mother-child behavior in a placebo-controlled trial of methylphenidate in preschoolers with ADHD and developmental delays. A similar paradigm was used in a placebo-controlled trial of naltrexone in children with ASD (Kolmen et al. 1995).

Direct observational measures have several potential advantages. First, they can be conducted blind to the intervention, which may be especially valuable in behavioral interventions that do not include a blinded placebo control group. Second, behaviors can be measured at the individual level. By definition, rating scales have a pre-selected set of behaviors that may not capture individual variation. Third, few rating scales are well suited to evaluate parent-child interaction, which may be a target of intervention. Finally, because direct observational methods may measure unique aspects of treatment outcome, they can serve as a valuable complement to parent and clinician-based measures.

Potential disadvantages of observational measures also warrant careful consideration. First, measures that assess behaviors in real time may require considerable training to achieve reliability (Arnold et al. 2000a). Failure to demonstrate reliability may provide a source of error due to raters. Second, behavioral observational

measures are constrained to capture a segment of time. Children with disruptive behavior often show variability in their symptoms. Observation on a “good day” may contribute to an exaggerated impression of improvement. Conversely, observation on a “bad day” may underestimate overall improvement. This variability may also contribute to error. Third, video recordings are vulnerable to technical failure and missing data. Finally, especially for video recordings, direct observational measures can be time-consuming and expensive. The time and expense of coding video recordings obviously goes up stepwise with the sample size.

In a trial comparing risperidone alone versus risperidone plus parent training, the Research Units on Pediatric Psychopharmacology – Autism Network (RUPP) included four types of outcome measures: standardized rating scales, blinded clinician ratings, blinded assessment of narratives based on parent-defined target symptoms, and direct observation of mother-child interactions (Arnold et al. 2003; Johnson et al. 2009; Scahill et al. 2009; Aman et al. 2009). The study found that subjects in the risperidone plus parent training group reported significantly lower scores on standardized ratings scales (e.g., assessing behaviors such as aggression, tantrums and self-injury) and blinded clinician ratings (Clinical Global Improvement Scale) than subjects on medication alone (Aman et al. 2009). The direct observation measure assessed both parent and child behavior and was based, in-part, upon the functional analysis literature (the results of the direct observation measure have not been published elsewhere). The use of functional behavioral analysis involves observing a child and parent in 3–4 different situations in order to determine the possible function of the behavior(s) of concern (Hanley et al. 2003; Iwata et al. 1982). Traditionally, these situations have included five conditions: (1) Alone (where the child is left alone and provided few if any materials), (2) Play (a control condition in which the child is provided moderately interesting toys and given intermittent attention by a parent), (3) Attention (in which the parent is instructed to immediately attend to the child whenever an agreed upon target behavior occurs), (4) Demand (in which demands are terminated whenever an agreed upon target behavior occurs), and (5) Tangible Restrictive (in which highly desirable items are removed from the child and returned whenever an agreed upon target behavior occurs). The quantitative outcome measures of interest are counts of maladaptive behavior, compliant behavior and parent behavior (e.g., use of praise, rapidly repeating demands).

The purpose of the present paper was to examine the ability of a structured direct observational measure (Standardized Observation Analogue Procedure [SOAP]) to detect change in both parent and child behaviors. There were three hypotheses of interest. First, we tested whether the SOAP would show improvement from baseline to endpoint with risperidone, a treatment of established efficacy in children with ASD and irritability. Second, we examined whether the SOAP would discriminate between treatments (medication alone versus medication plus parent training), especially on observations related to parental behaviors (a primary target of parent training). Finally, we predicted that change on SOAP measures of child maladaptive behavior would be associated with changes in the study’s primary outcome measures (the Home Situations Questionnaire and the Aberrant Behavior Checklist Irritability subscale). However, we predicted that changes on SOAP measures of parent behavior were unlikely to be associated with the study’s primary outcome measures, as these measures were focused on child behavior.

Methods

Sample Characteristics

Subjects were recruited at three study sites: The Ohio State University, Yale University and Indiana University. The institutional review boards at each site approved the investigation. Parents/legal guardians provided written consent to participate in the study, while children provided assent (when able). Inclusion criteria entailed the following: (a) diagnosis of ASD (autistic disorder, pervasive developmental disorder not otherwise specified [PDD-NOS], or Asperger's disorder) based upon clinical evaluation and corroborated by the Autism Diagnostic Interview—Revised (ADI—R)(Rutter et al. 2003); (b) 4 to 13 years of age, inclusive; (c) $IQ \geq 35$ and mental age ≥ 18 months (based upon either the Stanford Binet 5th Edition or Mullen–Mullen 1989; Roid 2003); (d) a raw score of ≥ 18 on the Irritability subscale (ABC-I) of the parent-rated Aberrant Behavior Checklist (ABC; Aman et al. 1985a); (e) a Clinical Global Impressions—Severity Score ≥ 4 (Arnold et al. 2000b); and (f) free of psychotropic medications for a minimum of two weeks (with the exception of anticonvulsant therapy for seizures).

Design

A 24-week, double-blind parallel groups design was employed to compare risperidone alone (MED) versus risperidone plus parent training (COMB). In order to improve our ability to interest families in study participation, a 3:2 randomization to COMB and MED was used. Subjects in COMB treatment received parent training from a behavior therapist trained on the RUPP parent manual (Johnson et al. 2007). All parent trainers were doctorally-prepared or masters-level clinicians. Parent training sessions (60–90 min in length) included 11 core sessions and up to 3 optional sessions delivered over 16-weeks. All sessions were conducted individually, allowing the behavior therapist to customize the curriculum to match the functioning level (e.g., verbal vs. nonverbal) and target symptoms (e.g., tantrums vs. aggression) of each child. Two face-to-face and one telephone booster session were provided between weeks 16 and 24. The training curriculum focused upon teaching antecedent-prevention strategies, the use of positive reinforcement, compliance training and instructional methods for teaching new skills. The curriculum included session scripts, activity sheets, video tapes providing examples of the different skills and homework assignments. Each family was provided with individualized homework assignments between sessions, and parents were taught to collect data on children's behavior. Training was further individualized by selecting optional sessions that met each family's needs, including topics such as toilet training, sleep problems, or implementing a token economy. Behavior therapists conducted two home visits—the first occurring during the initial couple of weeks of parent training to assess the home situation in terms of physical layout, safety issues, etc.; the second after 16 weeks of training to assess compliance with treatment recommendations (e.g., was there a visual schedule posted in the bathroom, was there a time out area in the living room).

Prior to initiating the study, therapists were certified by the parent training supervisors. Approval was based upon reviewing and rating a set of 11 taped pilot patient

sessions (in which a treatment fidelity level of 80 % was established). All parent training sessions were videotaped, and 10 % were randomly selected and reviewed for ongoing fidelity during the five years of the study.

Both the MED and COMB groups were placed on risperidone. The dose of risperidone was flexibly adjusted to a maximum of 1.75 mg/day for subjects weighing 14–20 kg, 2.5 mg/day for subjects weighing >20 and ≤45 kg and 3.5 mg/day for those weighing >45 kg. Subjects who failed to respond to risperidone (or who had significant side effects, even if the medication dose was decreased) at week 8 were switched to aripiprazole. [See Scahill et al. 2009 and Aman et al. 2009 for further details regarding dosing of aripiprazole]

Study Instruments and Dependent Measures

Standardized Observation Analogue Procedure (SOAP)

SOAP observations were conducted in a clinic room (each site had a slightly different sized room) containing a one-way observation mirror. Only the mother and child were in the clinic room while study staff observed from behind the mirror. Communication between parent and staff was maintained via a bug-in-the-ear device and all sessions were video-taped. A series of four, 10-minute mother-child interactions (conditions) were conducted once at baseline (Week 0, no medication) and once at study endpoint (Week 24, risperidone+parent training). The four conditions included: a) *Free Play Condition*—Based upon parent selection from a list of available toys, each child was provided an individualized set of 9 toys comprising three highly desirable, three moderately desirable, and three neutral toys. Parents were then instructed to interact and play with their children as they would typically do at home. Free Play served as a control condition: adult attention was given freely, preferred toys were available, and no demands were placed upon the child; b) *Social Attention Condition*—Neutral toys were placed in the room. Parents were provided with a questionnaire to complete as well as reading materials. They were instructed to focus on the materials and to respond to their child's request for attention as they would typically do at home when engaged in an important task (e.g., when caring for another child in the home, talking on the telephone); c) *Demand Condition*—The same neutral toys from the Social Attention Condition remained in the room. The parents were provided with a list of 20 standard demands (e.g., "Give me the book") and asked to select 10 demands that the child could do but might resist when asked to complete. Appropriate materials related to the demands were placed in the room, and parents were cued every minute (via the "bug in the ear" device) to issue one of the demands. The parents were told to do what they normally did at home when giving their child directions; and d) *Tangible Restriction Condition*—Five moderately-to-highly-desirable toys were selected from the toy list (as well as a favorite toy brought from home), and parents were asked to engage their child in play with one of the toys for about 60 s. Then, parents were asked (via the "bug in the ear" device) to take the toy from the child and attempt to manage the child's behavior (e.g., crying, demanding a return of the toy) just as they would at home. After about 30 s, the parent was allowed to re-involve the child with another toy, if the child had not already done so. This sequence of removing and re-introducing highly desirable toys was repeated five times.

Coding Procedures The middle five minutes (minutes 3–7) of each SOAP condition recorded on video was coded by blinded raters. A partial interval recording system was used, employing a 10-second observe, 5-second record system via the Procoder System DV software (Tapp 2003). The following six behaviors were coded:

Child Behaviors: a) Inappropriate Behavior included disruptive behaviors (e.g., climbing on furniture, running out of the room, turning out the lights), or more serious behaviors such as aggression, tantrums and self-injury. Inappropriate behavior was coded as present/absent across all four conditions. b) Compliance/Noncompliance: This category of behavior was only coded during the Demand condition. A child was scored as being compliant if a given request was followed within 60 s. The rationale for such a long time (versus within 10–20 s) was that some tasks involved a number of steps that could not be completed within a shorter period (e.g., “find all of the aces in a deck of cards”).

Parent Behaviors: a) Restrictive Comments: These involved statements directed towards the child with the use of the word(s), “No,” “Don’t,” and “Stop;” b) Positive Reinforcement: This category included praise as well as other forms of social reinforcement (e.g., hugs, “high-fives”); c) Repeated Demands (coded only in the Demand condition): This was the number of times that a parent repeated the same demand in an effort to get the child to comply; and d) Percent Contingent Reinforcement: This was defined as positive reinforcement within 20 s of child compliance (coded only during the Demand Condition).

Interobserver Reliability Agreement Procedures Prior to coding tapes, raters were trained to reliability with three training tapes. The project coordinator, who had over 10 years of experience conducting behavioral research, served as the trainer as well as the expert rater. The primary rater was an undergraduate psychology student. Reliability was determined by dividing the percent agreement (total agreements for each coded variable for each session) by the total agreements plus total disagreements times 100 for 10 % of randomly-selected sessions. Percent agreement ranged from 88 % to 100 % (mean) across the six behaviors coded during the SOAP.

Parent Ratings

As reported elsewhere, parents completed a set of standardized questionnaires at the time of each clinic visit. For the purpose of the present paper, the two primary outcome measures, the Home Situations Questionnaire (Barkley et al. 1999) and the Irritability subscale from the Aberrant Behavior Checklist (Aman et al. 1985a), were used.

Home Situations Questionnaire (HSQ; Barkley et al. 1999) The HSQ is a 25-item parent-rated scale that assesses noncompliant behavior in a variety of everyday situations (e.g., when at meals, when getting dressed). For situations where problems occurred, severity was scored on a 1–9 Likert scale (ranging from mild to very severe). Whereas the original HSQ contained only 20 items, five additional items felt to be especially pertinent to children with ASD were added for the present study (e.g., a change in the arrangement of a familiar setting). This modified version was initially used in our parent-training feasibility study (RUPP Autism Network 2007) and shows reliability and discriminate validity (Chowdhury et al. 2010). At study baseline, parents/ caregivers were asked to rate their children’s behavior over the past

two weeks. During the course of the study, ratings were based upon behavior since the previous clinic visit. A single HSQ score was calculated by adding all Likert scores and dividing by 25. Only the baseline and final, 24-week HSQ scores were used in the current analysis.

Aberrant Behavior Checklist (ABC; Aman et al. 1985a) The ABC is a 58-item, teacher or parent-completed rating scale comprising the following subscales (a) Irritability (15 items), (b) Lethargy/Social Withdrawal (16 items), (c) Stereotypic Behavior (7 items), (d) Hyperactivity/Noncompliance (16 items), and (e) Inappropriate Speech (4 items). Each item was rated using a four-point scale (from “not at all a problem” to “the problem is severe in degree”). The rating scale was empirically derived, has sound psychometric characteristics, and has been shown to be sensitive to treatment effects (Aman et al. 1985b). Parents/caregivers were asked to rate their child’s behavior using the identical time as used for the HSQ. Only the ABC-I scores from baseline and the 24-week visit (or end-point) were used in the current analysis.

Statistical Analysis

Subjects in the study were analyzed on the basis of the Intent-to-Treat principle. For the initial analysis of SOAP variables, a multiple comparison with the Tukey-Kramer test was made among four different conditions on Child Inappropriate Behavior. Changes in parent and child behaviors between baseline and the study endpoint, within the four SOAP conditions regardless of treatment groups, were evaluated with paired t-tests. Next, analysis of covariance (ANCOVA) adjusting for baseline values and sites was conducted to examine the treatment effects (MED vs. COMB) on changes in SOAP variables. Since there were many zeros on restrictive/positive raw scores in each condition, Poisson regression model adjusting over dispersion was used. Last, a correlation analysis was performed using *Pearson* correlation coefficients. All analyses were performed using SAS Version 9.2 (SAS Institute Inc, Cary, NC), with statistical significance set at *P* less than 0.05 using 2-sided tests.

Results

Subjects

A total of 199 potential participants were screened and 124 (62.3 %) were randomized to the study. Forty-nine (39.5 %) were assigned to Medication Only (MED) and 75 (60.5 %) to the Combined Treatment condition (COMB). Eighty-five percent of the sample ($N=105$) were male; the mean age for the MED group was 7.5 years ($SD=2.8$) and was 7.4 years ($SD=2.2$) for the COMB group. In the randomized group ($N=124$), 81 (65.3 %) were diagnosed with autistic disorder, 35 (28.2 %) with PDD-NOS, and 8 (6.5 %) with Asperger’s disorder. The two treatment groups were not significantly different for household income, parental education, or educational placement. Similarly, most clinical characteristics, including baseline HSQ and ABC-I scores, did not differ between treatment groups. However, the MED group had significantly lower scores on the Vineland Adaptive Behavior Scales and IQ test as well as having

been more likely to be prescribed anticonvulsants at baseline than subjects in the COMB group. Table 1 provides specific demographic information for the two treatment groups.

Exposure to Treatment

Subjects assigned to COMB received an average of 10.82 (SD=3.16) parent training (PT) sessions by week 24. Eighty-seven percent of subjects continued with risperidone until the end of the trial; the remaining 13 % of subjects were switched to aripiprazole by week 8 of the trial.

Outcome Measures

Improvement from Baseline to Endpoint on the SOAP

Prior to collapsing all Child Inappropriate Behavior into a single baseline score across the treatment groups, we examined scores across the four SOAP conditions. As would be expected, the lowest rate of Inappropriate Behavior (20 % of intervals) was noted in the Free Play Condition (parental attention was freely given, preferred toys were available, and few demands were made of the child). Significantly higher rates of Inappropriate Behavior were noted in the Social Attention Condition (32 % of intervals), Demand Condition (40 % of intervals) and Tangible Restriction Condition (42 % of intervals) ($p < .0001$). Because this trend was evident in both treatment

Table 1 Subject Demographics (N=124)

Variable	MED (%/SD)	COMB (%/SD)	p
Gender			.45
Male (105)	40 (81.6)	65 (86.7)	
Female (19)	9 (18.4)	10 (13.3)	
Age	7.50 (2.80)	7.38 (2.21)	.86
Race			.15
White	34 (69.4)	59 (78.7)	
Hispanic	7 (14.3)	4 (5.3)	
African American	7 (14.3)	9 (12.1)	
Other	1 (2.0)	3 (4.0)	
IQ Category			.02
Average			
Borderline			
Mild ID			
Moderate ID			
PDD Diagnosis			.83
Autistic Disorder			
PDD-NOS			
Asperger’s Disorder			

groups at baseline, we collapsed these scores into a single Child Inappropriate Behavior index across the treatment groups.

Table 2 provides the means (SDs) for parent and child measures across the four SOAP conditions (collapsed across the two treatment groups). No statistically significant decreases in Inappropriate Behavior were noted for either the Free Play or Social Attention conditions. Conversely, significant decreases were observed in the Demand (27.5 %, $p=0.0002$) and Tangible Restrictive conditions (21.4 %, $p=0.012$). Child compliance was unexpectedly high at baseline in the Demand Conditions (mean 75 % (SD 25 %)). Despite this, there was a small, but statistically significant increase in the rate of compliance (12 %, $p=0.004$) noted at follow-up. Along with increased compliance and decreased inappropriate behavior in the Demand Condition, changes in some parent variables were also noted at 24 weeks. For example, parents showed a significant increase in their use of *positive reinforcement* during both the free play ($p=0.004$) and demand conditions ($p=0.001$), used significantly fewer *restrictive statements* during the Social Attention condition ($p=0.03$), and showed a significant decrease in repeated demands during the Demand condition ($p<.0001$).

Table 2 Combined treatment groups

SOAP Variable (P = Parent; C = Child)	Baseline	Endpoint	P-value ^c
Free Play			
Inappropriate Mean % Intervals ^a (C)	20 (23)	17 (21)	0.17
Restrictive Mean Raw Score ^b (P)	0.88 (1.72)	1.10 (1.79)	0.27
Positive Mean Raw Score ^b (P)	1.34 (2.01)	2.23 (3.19)	0.004
Social Attention			
Inappropriate Mean % Intervals ^a (C)	32 (29)	29 (27)	0.41
Restrictive Mean Raw Score ^b (P)	1.45 (3.27)	0.65 (1.51)	0.03
Positive Raw Score ^b (P)	0.30 (0.83)	0.53 (1.24)	0.13
Demand			
Inappropriate Mean % Intervals ^a (C)	40 (27)	29 (21)	0.0002
Mean % Compliance ^a (C)	75 (25)	84 (19)	0.004
Restrictive Mean Raw Score ^b (P)	1.49 (2.05)	1.21 (2.06)	0.39
Positive Mean Raw Score ^b (P)	1.48 (1.99)	2.42 (2.62)	0.001
Repeated Mean Raw Score ^a (P)	10.25 (7.15)	7.33 (6.61)	<.0001
Mean % Contingent Reinforcement ^a (P)	39 (29)	41 (25)	0.77
Tangible Restriction			
Inappropriate Mean % Intervals ^a (C)	42 (27)	33 (24)	0.012
Restrictive Mean Raw Score ^b (P)	2.32 (3.30)	1.62 (2.18)	0.10
Positive Mean Raw Score ^b (P)	1.13 (1.97)	1.58 (2.33)	0.09

^a Based upon % of intervals behavior observed (30 ten-second intervals)

^b Based upon number intervals behavior was observed (out of a possible 30)

^c P-value from *paired t-test*

A correlation analysis was conducted to examine the relationship between changes on SOAP measures and changes in the study's two primary outcome measures (the HSQ and ABC-I). No statistically significant correlations were found.

MED Versus COMB (Additive Effect of Parent Training)

As shown in Table 3, no differential treatment effects for MED vs. COMB were evident on measures of Child Inappropriate Behavior. However, parents in the COMB group used significantly fewer restrictive statements in the Social Attention ($p=0.001$) and Tangible Restriction ($p=0.046$) conditions and provided significantly more reinforcement (contingent on child compliance) ($p=0.01$) than parents in the MED group.

Discussion

SOAP measures were able to detect treatment change over time. Yet, improvement on SOAP measures (for both child and parent variables) was not found to correlate with the study's primary outcome measures (the HSQ and ABC-I). This suggests that direct observation of mother-child interactions may be assessing something different than the standardized questionnaires. To examine if this finding was consistent with prior research, we identified a total of 14 studies in which both the ABC and direct observations were used. While both types of assessments were often shown to be sensitive to behavior change in response to treatment, rarely was a correlation of the two types of measures obtained. Only in the original ABC psychometric study (Aman, et al. 1985b) was the ABC found to correlate with direct observations. However, in this case the study did not involve treatment outcome. To determine if group differences on IQ might have impacted the study findings, the possible influence of IQ was evaluated in our initial paper (Aman, et al. 2009). Using the study's two primary outcome measures (HSQ and ABC-I), no interaction of treatment and treatment-by-time effects were found with IQ.

The SOAP's ability to detect differential treatment effects for the two groups was limited to changes in some parental behaviors. As these are the types of parent skills/behaviors that were included in the PT curriculum, it is likely that these changes were the result of the combined treatment (and might be difficult to detect on other measures such as the CGI or behavior rating scales). Consistent with prior research (e.g., Barkley et al. 1984), the Demand Condition appeared to be the most sensitive to detecting changes in both child and parent behavior and was second to Tangible Restriction in having the highest frequency of baseline child inappropriate behavior. In the Free Play condition (where parental attention was provided and few demands were placed upon the child), significantly fewer behavior problems occurred.

The main challenge in using in-clinic direct observation measures was our ability to successfully mimic a situation where the child (and parent) would exhibit behavior similar to what is seen in the home. Children with ASD can respond to novel situations in different ways, from being quiet and withdrawn to engaging in more challenging behavior. While 26.6 % of subjects failed to display any inappropriate

Table 3 Between treatment groups (expressed as mean % or raw score)

SOAP Variable (P = Parent; C = Child)	MED Only		COMBINED		Adjusted Mean Difference ^d (SE)	p-value ^e	Effect size ^f	Effect size ^g
	Baseline	Endpoint	Baseline	Endpoint				
Free Play								
Inappropriate % ^a (C)	23 (25)	20 (20)	19 (23)	15 (21)	-0.05 (0.04)	0.24	0.21	0.24
Restrictive Raw Score ^b (P)	1.06 (2.10)	1.12 (1.79)	0.78 (1.44)	1.09 (1.81)	0.15 (0.30)	0.62	0.09	0.08
Positive Raw Score ^b (P)	1.67 (2.48)	2.15 (3.51)	1.14 (1.64)	2.28 (3.01)	0.26 (0.29)	0.36	0.13	0.38
Social Attention								
Inappropriate % ^a (C)	31 (30)	33 (30)	33 (29)	27 (25)	-0.07 (0.06)	0.20	0.24	0.26
Restrictive Raw Score ^b (P)	1.25 (1.71)	1.12 (1.98)	1.51 (3.53)	0.37 (1.06)	-1.20 (0.37)	0.001	0.40	0.82
Positive Raw Score ^b (P)	0.22 (0.48)	0.32 (1.17)	0.35 (0.99)	0.65 (1.27)	0.63 (0.46)	0.15	0.76	0.51
Demand								
Inappropriate % ^a (C)	41 (30)	32 (24)	40 (26)	27 (20)	-0.05 (0.05)	0.25	0.18	0.23
% Compliance ^a (C)	78 (23)	81 (22)	74 (26)	86 (18)	0.04 (0.04)	0.39	0.16	0.20
Restrictive Raw Score ^b (P)	1.47 (1.73)	1.41 (2.23)	1.51 (2.23)	1.09 (1.97)	-0.41 (0.30)	0.17	0.20	0.22
Positive Raw Score ^b (P)	1.56 (2.02)	1.94 (2.36)	1.42 (1.94)	2.70 (2.76)	0.34 (0.24)	0.15	0.17	0.13
Repeated Raw Score ^b (P)	10.44 (8.41)	8.38 (7.88)	10.11 (6.34)	6.61 (5.71)	-1.12 (1.28)	0.38	0.16	0.17
% Contingent Reinforcement ^a (P)	43 (31)	34 (19)	37 (27)	46 (30)	0.17 (0.06)	0.01	0.59	0.64

Table 3 (continued)

SOAP Variable (P = Parent; C = Child)	MED Only		COMBINED		Adjusted Mean Difference ^d (SE)	p-value ^e	Effect size ^f	Effect size ^g
	Baseline	Endpoint	Baseline	Endpoint				
Tangible Restriction								
Inappropriate % ^a (C)	40 (24)	37 (25)	43 (29)	31 (24)	-0.07 (0.05)	0.19	0.26	0.29
Restrictive Raw Score ^b (P)	1.88 (2.71)	2.15 (2.92)	2.56 (3.74)	1.29 (1.53)	-0.57 (0.28)	0.046	0.17	0.27
Positive Raw Score ^b (P)	1.06 (1.70)	1.21 (1.76)	1.18 (2.13)	1.80 (2.60)	0.33 (0.22)	0.30	0.17	0.14

SOAP Standardized Observation Analogue Procedure; SE standard error

^a Based upon % of intervals behavior observed (30 ten-second intervals)

^b Based upon number intervals behavior was observed (out of a possible 30)

^c Data are presented as (unadjusted) Mean (SD)

^d Adjusted Mean Difference means a difference of Least Squares Means from ANCOVA model controlling for Baseline score and site

^e P-value from ANCOVA model adjusting for Baseline score and site. Bold indicates p-value less than 0.05. Bold indicates P-value less than 0.05

^f Effect size calculated from adjusted mean difference divided by pooled standard deviation at baseline

^g Effect size calculated from adjusted mean difference divided by pooled standard deviation at endpoint

behavior during the Free Play condition, only 3.3 % failed to do so in the Demand condition and 1.1 % in the Tangible Restriction condition. Consequently, analogue mother-child interactions did provide an opportunity to observe maladaptive behavior within a clinic setting. Despite this, the overall rates of compliance during the Demand condition (arguably the most important observation if working with a child whose behavior tends to function as a means of escaping demands) were unexpectedly high at baseline (around 75 %). Consequently, a number of modifications in the SOAP protocol may need to be considered if it is to provide a more accurate, valid, and sensitive assessment of child behavior.

One option is to conduct shorter, but more frequent observations, limited to a possible free play “warm-up” basal session, followed by a set of parental demands. This may be a more efficient way to use direct observations in clinic settings. The use of only this single condition would allow for repeated observations in order to obtain a greater number of pre- and post-treatment data points. One potential problem with this approach is that not every child’s inappropriate behavior serves an “escape from demand” function. For example, some children might engage in maladaptive behavior in order to obtain parental attention. Another possible way to lower the baseline rate of compliance may be to add more (or only) preferred toys (rather than including moderately preferred toys). This might make it more difficult for children with ASD to transition easily from play to a demand situation. A third possible adjustment might be to spend more time identifying demands that children are capable of doing but are likely to refuse to do during the observation. Adding items that require more effort or more sustained time on task may help to achieve this. Finally, it may also be helpful to have a number of baseline assessments to address child reactivity so that both members of the diad are truer to their usual roles.

Another option for decreasing baseline compliance rates is to require compliance within a shorter time period (e.g., a single 10-second interval). Because many of the subjects in this study were higher functioning, many demands required multiple steps (e.g., complete five math problems). Since compliance was determined when the task was completed, enough time needed to be allotted for such multi-step tasks to be done by the child. However, revising the coding system to one where initiation of compliance is coded (versus task completion) would allow for a much shorter time requirement (e.g., 10 s) and a lower baseline compliance rate.

The SOAPs general inability to detect group effects may have been due in part to the drug’s large effect size (see Aman et al. 2009), providing an artificially high ceiling for detecting the additive effects of parent training. Despite this, the trend showed a greater drop in inappropriate behavior in the COMB group than in MED alone. However, the extremely large variance in many of the SOAP categories resulted in such differences not being statistically significant. The finding of significant improvements in some parent behaviors for families assigned to COMB suggests that PT led to positive changes in parenting behaviors.

Finally, the lack of a positive correlation between SOAP measures and the two primary outcome measures may not be that surprising. It suggests that direct observations of mother-child interactions may be assessing something very different from parental global ratings of a child’s behavior. If so, this may provide an even stronger case for the need to conduct direct observations in addition to collecting caregiver ratings within psychopharmacology trials. Caregiver ratings provide an average

estimate for the rating window (e.g., week or month) without the stress of a clinic visit. On the other hand, the structured observation in the clinic functions like an exam in an academic setting: it tests systematically and stringently for the acquired skill or knowledge “on demand.” Taken together, they provide a comprehensive evaluation of progress. Structured observations, while commonly used among clinicians with training in applied behavior analysis, is rarely used in outpatient clinics to document response to medication. However, most prescribers are attentive to both parent and child behavior during office visits (noticing activity level, tantrums and parent limit-setting) and often use such observations in considering treatment options. The use of a short, 5-to-10 min standardized parent-child observation may provide a more structured and consistent means of interpreting behavior seen in the clinic. Perhaps a clinician-friendly protocol could be developed in which the parent instructs the child to do something while the parent fills out the scales, and 5 min of that activity is observed with a check-sheet.

Acknowledgements We wish to acknowledge Catherine A. Belasco, for her valuable assistance in the coding of direct observation data and preliminary data analysis. We also thank all the families who participated in this study.

The views expressed in this article are those of the authors and do not necessarily reflect the official position of the National Institute of Mental Health, the National Institutes of Health, or any other part of the U.S. Department of Health and Human Services. NIMH encourages publication of results and free scientific access to data.

Disclosures Dr. Handen has received research support from Eli Lilly, Curemark and Bristol-Myers Squibb; Dr. Scahill has served as a consultant or has received research support from Boehringer-Ingelheim, NeuroSearch, Pfizer, Hoffman, BioMarin, Shire and Seaside Therapeutics. Dr. Aman has served as a consultant or received research support from Biomarin Pharmaceuticals, Confluence Pharmaceutica, Forest Research, Hoffman LaRoche, ProPhase and Supernus Pharmaceutica. Dr. Arnold has received research funding or advisory board honoraria from AstraZeneca, Biomarin, Curemark, Eli Lilly, Novartis, Noven, Seaside Therapeutics and Shire and travel support from Noven. Dr. Stigler has received research support from Bristol-Myers Squibb, Curemark, Eli Lilly, Forest Research Institute, Janssen, Novartis and Seaside Therapeutics.

References

- Abikoff, H., Hechtman, L., Klein, R. G., Gallacher, R., Fleiss, K., Etcovitch, J., et al. (2004). Social functioning in children with ADHD treated with long-term methylphenidate and multimodal psychosocial treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 820–829.
- Ahearn, W. (2002). Effect of two methods of introducing foods during feeding treatment on acceptance of previously rejected items. *Behavioral Interventions*, *17*, 111–127.
- Aman, M. G., McDougle, C. J., Scahill, L., Handen, B. L., Arnold, L. E., Johnson, C. J., et al. (2009). Medication and parent training in children with pervasive developmental disorders and serious behavior problems. *Journal of the American Academy of Child and Adolescent Psychiatry*, *48*, 1143–1154.
- Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. J. (1985a). The Aberrant Behavior Checklist: a behavior rating scale for the assessment of treatment effects. *American Journal of Mental Deficiency*, *89*, 485–491.
- Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. J. (1985b). Psychometric characteristics of the Aberrant Behavior Checklist. *American Journal of Mental Deficiency*, *89*, 492–502.
- Arnold, L. E., Aman, M. G., Martin, A., Collier-Crespin, A., Vitiello, B., Tierney, E., et al. (2000a). Assessment in multisite randomized clinical trials of patients with autistic disorder: the Autism RUPP Network. *Journal of Autism and Developmental Disorders*, *30*, 99–111.

- Arnold, L. E., Aman, M. G., Martin, A., Collier-Crespin, A., Vitiello, B., Tierney, E., Asarnow, R., Bell-Bradshaw, F., Freeman, B. J., Gates-Ulanet, P., Klin, A., McCracken, J. T., McDougle, C. J., McGough, J. J., Posey, D. J., Scahill, L., Swiezy, N. B., Ritz, L., & Volkmar, F. (2000b). Assessment in multisite randomized clinical trials (RCTs) of patients with autistic disorder: the Autism RUPP network. *Journal of Autism and Developmental Disorders*, *30*, 99–111.
- Arnold, L. E., Vitiello, B., McDougle, C., Scahill, L., Shah, B., Gonzalez, N. M., et al. (2003). Parent-defined target symptoms respond to risperidone in RUPP Autism Study: customer approach to clinical trials. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 1443–1450.
- Barkley, R. A., & Cunningham, C. E. (1979). The effects of methylphenidate on the mother-child interactions of hyperactive children. *Archives of General Psychiatry*, *36*, 201–208.
- Barkley, R. A., Edwards, G. H., & Robin, A. L. (1999). *Defiant teens: A clinician's manual for assessment and intervention* (pp. 196–198). New York: Guilford Press.
- Barkley, R. A., Karlsson, J., Strzelecki, E., & Murphy, J. V. (1984). Effects of age and Ritalin dosage on the mother-child interactions of hyperactive children. *Journal of Consulting and Clinical Psychology*, *52*, 750–758.
- Chowdhury, M., Aman, M. G., Scahill, L., Swiezy, N. B., Arnold, L. E., Lecavalier, L., et al. (2010). The Home Situations Questionnaire–PDD Version: factor structure and psychometric properties. *Journal of Intellectual Disabilities Research*, *54*, 281–291.
- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle: Pearson/Merrill Prentice Hall.
- Cunningham, C. E., & Barkley, R. A. (1979). The interactions of normal and hyperactive children with their mothers in free play and structured tasks. *Child Development*, *50*, 217–224.
- Handen, B. L., Feldman, H., Lurier, M., & Husar, P. J. (1999). Efficacy of methylphenidate among preschoolers with developmental disabilities and ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*, 805–812.
- Hanley, G. P., Iwata, B. A., & McCord, B. (2003). Functional analysis of problem behavior: a review. *Journal of Applied Behavior Analysis*, *36*, 147–185.
- Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1982). Toward a functional analysis of self-injury. *Analysis and Intervention in Developmental Disabilities*, *2*, 3–20.
- Johnson, C. R., Butter, E. M., Handen, B. L., Sukhodolsky, D. G., Mulick, J., Lecavalier, L., et al. (2009). Standardized observation analogue procedure (SOAP) for assessing parent and child behaviors in clinical trials. *Journal of Intellectual & Developmental Disability*, *34*, 230–238.
- Johnson, C. R., Handen, B. L., Butter, E., Wagner, A., Mulick, J., Sukhodolsky, D. G., et al. (2007). Development of a parent training program for children with pervasive developmental disorders. *Behavior Interventions*, *22*, 201–221.
- Kahng, S. W., Leak, J. M., Vu, C., & Mishler, B. (2008). Mechanical restraints as positive reinforcers for aggression. *Behavioral Intervention*, *23*, 137–142.
- King, B. H., Hollander, E., Sikich, L., McCracken, J. T., Scahill, L., Bregman, J. D., et al. (2009). Lack of efficacy of citalopram in children with autism spectrum disorders and high levels of repetitive behavior. Citalopram ineffective in children with autism. *Archives of General Psychiatry*, *66*, 583–590.
- Kolmen, B. K., Feldman, H., Handen, B. L., & Janosky, J. E. (1995). Naltrexone in young autistic children: a double blind placebo-controlled crossover study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *34*, 223–231.
- McMahon, R. J., & Forehand, R. L. (2003). *Helping the noncompliant child: Family-based treatment for oppositional behavior* (2nd ed.). New York: Guilford Press.
- MTA Cooperative Group. (1999). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1073–1086.
- Mullen, E. J. (1989). *Mullen scales of early learning*. Bloomington: Pearson Assessments.
- Pelham, W. E., Hoza, B., Pillow, D. R., Gnagy, E. M., Kipp, H. L., Greiner, A. R., et al. (2002). Effects of methylphenidate and expectancy on children with ADHD: behavior, academic performance, and attributions in a summer treatment program and regular classroom settings. *Journal of Consulting and Clinical Psychology*, *70*, 320–335.
- Research Units on Pediatric Psychopharmacology Autism Network. (2002). A double-blind, placebo-controlled trial of risperidone in children with autistic disorder. *The New England Journal of Medicine*, *347*, 314–321.
- Research Units on Pediatric Psychopharmacology Autism Network. (2005). A randomized, double-blind, placebo-controlled, crossover trial of methylphenidate in children with hyperactivity associated with pervasive developmental disorders. *Archives of General Psychiatry*, *62*, 1266–1274.

- Roid, G. H. (2003). *Stanford-Binet intelligence scales* (5th ed.). Itasca: Riverside Publishing.
- RUPP Autism Network. (2007). Research Units on Pediatric Psychopharmacology [RUPP] Autism Network: feasibility trial of parent management training for children with pervasive developmental disorders. *Behavioral Interventions*, 22, 179–199.
- Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism diagnostic interview-revised*. Los Angeles: Western Psychological Services.
- Scahill, L., Aman, M., McDougle, C. J., Arnold, L. E., McCracken, J. T., Handen, B., et al. (2009). Trial design challenges when combining medication and parent training in children with pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 39, 720–729.
- Swanson, J. M., Wigal, S., Greenhill, L. L., Browne, R., Waslik, B., Lerner, M., et al. (1998). Analogue classroom assessment of Adderall in children with ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37, 519–526.
- Tapp, J. (2003). *ProCoder for digital video: User manual*. Nashville: John F. Kennedy Center, Vanderbilt University.
- Taylor, B., & Hoch, H. (2008). Teaching children with autism to respond to and initiate bids for joint attention. *Journal of Applied Behavior Analysis*, 41, 377–391.
- Webster-Stratton, C. H. (1996). Early intervention with videotape modeling: Programs for families of children with oppositional defiant disorder or conduct disorder. In E. D. Hibbs & P. S. Jensen (Eds.), *Psychosocial treatments for child and adolescent disorders: Empirically based strategies for clinical practice* (pp. 435–474). Washington: American Psychological Association.
- Yang, L. (2003). Combination of extinction and protective measures in the treatment of severely self-injurious behavior. *Behavioral Interventions*, 18, 109–121.